UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Single-molecule imaging of transcription dynamics

Alda Filipa Queirós Oliveira e Silva

Mestrado em Biologia Molecular e Genética

Dissertação orientada por:
Doutor Robert Manfred Martin
Professora Doutora Margarida Gama Carvalho

2019

Agradecimentos

Este projeto foi realizado no laboratório SAlmeida, no Institutito de Medicina Molecular e resultou de um ano de trabalho, esforço e dedicação, que me permitiu crescer a nível profissional, científico e pessoal. No entanto, isto não teria sido possível sem a contribuição de várias pessoas.

Em primeiro lugar, queria agradecer ao meu orientador externo, Doutor Robert Martin, pela confiança que depositou em mim para trabalhar neste projeto, por todo o conhecimento que me transmitiu, pelo apoio, disponibilidade, compreensão, paciência e dedicação. Agradeço também ao Doutor Sérgio Almeida por me ter recebido no seu laboratório e me ter dado a oportunidade de me iniciar no mundo da investigação científica. Agradeço-lhe também a sua ajuda, o incentivo ao espírito crítico, a compreensão, a disponibilidade e a sua ajuda preciosa na resolução dos problemas que foram aparecendo. Gostava também de agradecer à minha orientadora interna, Doutora Margarida Carvalho, pelo seu acompanhamento, disponibilidade e ajuda ao longo do projeto. Um agradecimento ao Instituto de Medicina Molecular por me ter recebido e dado os meios necessários para a realização deste projeto.

Um enorme obrigado a todos os meus colegas de laboratório pela ajuda essencial para a realização deste projeto, pela paciência que tiveram em ensinar e corrigir uma cabeça-no-ar e pelo apoio em todos os momentos. Obrigada à Madalena e ao João pela constante boa disposição. Obrigada à Cristiana e à Inês pela vossa calma e sensatez. Obrigada ao Ram e à Naike por todo o conhecimento e incentivo. Agradeço também aos elementos do Bioimaging e do Claus Azzalin Lab pela ajuda essencial na realização deste projeto.

Obrigada a todos os meus colegas de mestrado, principalmente à Bia, à Cláudia, ao Pedro, à Maria, à Sofia e à Isa pela sua amizade e ajuda nesta etapa da minha vida.

Apesar deste projeto ter sido desenvolvido em Lisboa, a minha terra natal é o Porto e lá também tenho amigos fantásticos que celebraram comigo todas as minhas conquistas e me apoiaram nas minhas dificuldades. Um grande obrigada a todos, em especial à Helena, à Fatinha, ao Fernando e à Raquel que acompanharam mais de perto o meu trabalho.

Um obrigada muito especial à minha família. Aos meus pais e à minha irmã por terem acompanhado sempre e apoiado de forma incondicional. Todos os sucessos que tive até agora e os que possa vir a ter, é a eles que devo. Às minhas tias São e Bela por todos os conselhos e por terem sido um exemplo para mim.

Quero também agradecer a uma pessoa muito especial, o Daniel, pelo carinho, paciência, apoio e dedicação. Foi incansável e acompanhou-me sempre, esteve ao meu lado nos bons e nos maus momentos. Obrigada por tudo.

Por fim, agradeço a todas as pessoas que contribuíram para o sucesso deste trabalho e que, apesar de não estarem mencionadas, foram fundamentais.

Abstract

Transcription is the molecular process that synthesizes an RNA molecule that is complementary to the DNA template. In eukaryotes, transcription is catalyzed by one of three RNA polymerases that share structural features and subunits but transcribe different types of genes. RNA polymerase II (RNA Pol II) transcribes genes that code for proteins and several long non-coding genes. RNA Pol II transcription involves three main stages: initiation; transition into productive elongation; and termination. This thesis focuses on transcription termination and aims at studying its functional links with RNA processing and the formation of non-canonical nucleic acids structures, such as R-loops. R-loops are by-products of transcription that also have been found to play a role in transcription regulation, however, it is not well understood how they affect transcription termination and mRNA processing. To decipher how R-loop formation affect the process of transcription and RNA processing, we directly examined with single-molecule sensitivity the synthesis of transcripts upstream and downstream the cleavage and polyadenylation site (CPAS) in the nucleus of living human cells and calculated the transcription rate of RNA Pol II for that regions. By using two different RNA labeling methods, PP7 and λN_{22}, we show that R-loops suppression impairs transcription termination. Further, we show that R-loop formation after the CPAS is not essential for an immediate transcription termination. A model for transcription termination, the torpedo model, suggests that exoribonuclease 2 (XRN2) digests the unprotected 5' end of the RNA transcript until it collides with the RNA Pol II, leading to its dissociation from the template DNA and favoring transcription termination. Here, we provide evidence for the torpedo model by observing a non-immediate termination when XRN2 is depleted. Moreover, we show that splicing inhibition impairs transcript cleavage and transcription termination. Our data provide new insights in the role of R-loops in transcription termination and mRNA processing.

Keywords: R-loops, transcription termination, splicing, single-molecule live cell imaging, RNA Pol II kinetics, torpedo model.

Resumo

A transcrição pela RNA polimerase II consiste principalmente na síntese de uma molécula de RNA complementar a um modelo de DNA. Este processo é fortemente regulado pela ligação de fatores de transcrição à polimerase, modificações epigenéticas ou formação de estruturas secundárias. Os R-loops são estruturas que resultam da hibridação do RNA produzido pela RNA polimerase II com a cadeia de DNA complementar. Estudos anteriores demonstraram que os R-loops têm um papel na regulação da transcrição, no entanto, ainda não se conhece de que forma afetam a terminação da transcrição e o processamento do mRNA. Neste projeto, analisámos os parâmetros cinéticos do processamento da extremidade 3' do mRNA e da terminação da transcrição e como a formação de R-loops afeta estes processos e o splicing. Para atingir os objetivos propostos, utilizamos microscopia confocal de células vivas com sensibilidade para detetar moléculas individuais, que permite acompanhar a produção e libertação/degradação de transcritos únicos em tempo real em células vivas e estudar os parâmetros cinéticos da RNA polimerase II durante a transcrição. Para analisar o comportamento cinético da RNA polimerase II na terminação da transcrição, foram utilizados genes repórter constituídos por dois exões separados por um intrão, provenientes do gene imunoglobulina M de murganho. No exão II foi inserida uma sequência de 24 stem-loops reconhecidos pela proteína de revestimento do bacteriófago PP7 e após o local de clivagem e poliadenilação (CPAS) foram inseridos 25 BoxB stem-loops, os quais são ligados pela proteína N_{22} do bacteriófago λ. Depois de induzir a transcrição do gene repórter com doxiciclina, as proteínas específicas ligam-se aos stem-loops presentes nos transcritos e, por microscopia confocal de células vivas com sensibilidade para detetar moléculas individuais, é detetado um ponto fluorescente correspondente ao local da transcrição.

Primeiro, realizamos uma descrição cinética do processamento da extremidade 3' do pré-mRNA e da terminação da transcrição. Observamos que, durante a transcrição do gene repórter IgM-1.7k-PY, a sequência de PP7 stem-loops no exão II é transcrita a uma taxa de cerca de 4,94 kbp/min, mas a transcrição dos BoxB stem-loops depois do CPAS não é detetada. Por isso, concluímos que a RNA polimerase II liberta-se do DNA antes de transcrever a sequência 25xBoxB stem-loops ou que há degradação do transcrito pós-CPAS, provavelmente pela exorribonuclease 2 (XRN2), ou seja, a terminação da transcrição é imediata. Também observamos que a clivagem do transcrito pré-CPAS acontece 35 ± 10 s após o início da transcrição da sequência dos PP7 stem-loops. Quando a concentração de doxiciclina é superior a 0,08 μg/ml, ocorre a síntese de vários transcritos em simultâneo, sendo possível detetar a produção de transcritos pré e pós-CPAS. Isto deve-se à maior densidade de polimerases no gene repórter, o que influencia o local e o momento da terminação da transcrição. Desta forma, há um atraso na libertação da RNA polimerase II do DNA modelo e transcrição da sequência de BoxB stem-loops.

Uma vez que no gene repórter IgM-1.7k-PY a sequência após o CPAS é propensa à formação de R-loops, testamos qual seria a influência da RNaseH1 na eficiência da terminação da transcrição. Verificamos que em 27% das células transfetadas com RNaseH1, havia produção de transcritos após o CPAS, o que significa que a terminação da transcrição não é imediata. Analisando a velocidade da RNA polimerase II, notamos que esta transcrevia mais rapidamente (6,82 kbp/min) a sequência dos PP7 stem-loops quando era detetada transcrição para além do CPAS. Uma possível explicação seria a degradação pela RNaseH1 de R-loops formados ao longo do gene que ajudam a controlar a velocidade da RNA polimerase II, levando a que a enzima não seja capaz de libertar-se do modelo de DNA antes de transcrever os BoxB stem-loops.

Para determinar como é que a formação de R-loops após o CPAS influencia a terminação da transcrição, foram utilizados dois genes repórter: um com uma sequência propensa à formação de R-loops após o CPAS (IgM-1.7k-PY-pA-baRFS) e outro com uma sequência que não forma R-loops (IgM-1.7k-PY-pY-pA-NRFS) no mesmo local. Para ambos os genes repórter apenas foi observada a síntese de transcritos pré-CPAS, indicando que a formação de R-loops após o CPAS não é determinante para a

terminação da transcrição como acontece noutros genes, por exemplo, o *SNRPN*. A nível cinético os genes repórter IgM-1.7k-PY-pA-baRFS e IgM-1.7k-PY-pY-pA-NRFS também tiveram resultados semelhantes ao IgM-1.7k-PY.

Existem dois modelos para a terminação da transcrição de RNA polimerase II: o modelo alostérico e o modelo do torpedo. O primeiro defende que o RNA polimerase II sofre uma alteração conformacional, após a transcrição do CPAS, libertando-se do modelo de DNA, enquanto o segundo postula que uma exorribonuclease (XRN2 em mamíferos) degrada o transcrito sintetizado após o CPAS e liberta a RNA polimerase II do DNA modelo quando a encontra. Questionamos qual o modelo de terminação da transcrição poderia explicar o processo de terminação no gene repórter IgM-1.7k-PY e para isso reduzimos a expressão da XRN2 e avaliamos o efeito na terminação da transcrição. Detetámos que, em 33% das células observadas, ocorria transcrição para além do CPAS, o que significa que a XRN2 é importante para uma terminação da transcrição imediata, constituindo assim uma evidência do modelo do torpedo.

Uma vez que o splicing e a terminação da transcrição são processos que estão interligados, averiguamos se um splicing ineficiente afetava a terminação da transcrição, utilizando o gene repórter IgM-1.7k-PYwsj, com um sinal de splicing fraco na extremidade 3' do intrão. Detetamos em todas as células observadas a transcrição para além do CPAS e um atraso na clivagem do pré-mRNA, relativamente ao gene repórter IgM-1.7k-PY. Além disso, notámos que a velocidade da RNA polimerase II era inferior durante a transcrição (3.71 kbp/min). Logo, a retenção do intrão afeta negativamente a clivagem e a terminação da transcrição, o que pode ser explicado pelo facto de o spliceossoma continuar ligado à RNA polimerase II, impedindo a ligação dos fatores necessários à clivagem do transcrito e à terminação da transcrição.

Uma vez que os R-loops têm um papel na regulação da transcrição, investigámos qual a influência da formação de R-loops antes do sinal de splicing na extremidade 3' do intrão no splicing, na clivagem e na terminação da transcrição, contruindo o gene repórter IgM-1.7k-3'intron-baRFS-PY. Este gene repórter possui com uma sequência propensa à formação de R-loops antes do sinal de splicing na extremidade 3' do intrão. Por qPCR, demostramos que 91% dos transcritos do IgM-1.7k-PY sofriam splicing, enquanto 98% dos transcritos do IgM-1.7k-PYwsj não. Surpreendentemente, 83% dos transcritos sintetizados a partir do gene repórter IgM-1.7k-3'intron-baRFS-PY sofriam splicing, indicando uma recuperação da eficiência do splicing através da formação de R-loops antes do sinal de splicing fraco na extremidade 3' do intrão. Por microscopia, só foi detetada transcrição do gene repórter IgM-1.7k-3'intron-baRFS-PY para além do CPAS em 18% das células observadas. Assim, mostramos que a formação de R-loops antes do PY tract pode restaurar parcialmente a eficiência do splicing, da clivagem e da terminação da transcrição, provavelmente, causando um abrandamento da polimerase e dando mais tempo para o reconhecimento do sinal de splicing pelo spliceossoma, aumentando assim a eficiência do splicing.

Em suma, estes resultados demostram que, quando a RNaseH1 é sobrexpressa, a velocidade de transcrição da RNA polimerase II é anormalmente elevada, ocorrendo transcrição para além do CPAS e que a formação de R-loops após o CPAS não essencial para a eficiência da clivagem e terminação da transcrição. Através da diminuição da expressão apresentamos evidências que confirmam o modelo do torpedo para a terminação da transcrição. Na presença de um sinal de splicing fraco na extremidade 3' do intrão, há retenção do intrão no transcrito e a clivagem do pré-mRNA e a terminação da transcrição são ineficientes. A inserção de uma sequência propensa à formação de R-loops antes desse sinal fraco de splicing, leva à recuperação parcial do splicing, clivagem e terminação da transcrição. Desta forma, os nossos resultados contribuíram para a compreensão do papel dos R-loops na terminação da transcrição e do processamento do mRNA.

Palavras-chave: R-loops, terminação da transcrição, splicing, microscopia confocal de células vivas com sensibilidade de para detetar moléculas individuais, cinética da RNA polimerase II, modelo do torpedo.

Table of contents

List of Tables

List of Figures

List of Abbreviations

3D – three dimensional

3'UTR – 3' untranslated region

λN_{22} – the 22 amino acids from the binding domain of bacteriophage λ antiterminator protein N

C9orf72 HRE – hexonucleotide repeat expanded in chromosome 9 open reading frame 72

ChIP – chromatin immunoprecipitation

CMV – cytomegalovirus

CPAS – cleavage and polyadenylation signal

CSR – class switch recombination

CTD – C-terminal domain

dsDNA – double stranded DNA

dsRNA – double stranded RNA

Dox – doxycycline

EMCCD – electron-multiplying charge-coupled device

FRT – flipase recombination target

Igs – immunoglobulins

IP – immunoprecipitation

iRFP – near-infrared fluorescent protein iRFP713

lncRNA – long non-coding RNA

mRNA – messenger RNA

NET-Seq – native elongating transcript sequencing

NRFS – non-R-loop forming sequence

Poly(A) – poly(A) tail

PSI – percent spliced in

PY – polypyrimidine

RFS – R-loop forming sequence

RNAi – RNA interference

RNA Pol II – RNA polymerase II

S – switch

SDS-PAGE – sodium dodecyl sulfate - polyacrylamide gel electrophoresis

SL – stem-loops

snRNAs – small nuclear RNAs

snRNP – small nuclear ribonucleoprotein

ssDNA – single stranded DNA

TA – triamcinolone acetonide

TC-NER – transcription-coupled nucleotide excision repair

TetO – tetracycline-resistance operon

TFI – total fluorescence intensity

TRC – transcription-replication conflict

TS – transcription site

TSS – Transcription start site

1. Introduction

1.1 Transcription

Transcription is one of the essential molecular processes in the cell that is the synthesis of an RNA molecule complementary to a DNA template. Transcription is catalyzed by RNA polymerases, enzymes responsible for the formation of phosphodiester bonds between ribonucleotides[1]. In eukaryotes, there are three types of RNA polymerases, in contrast to a single enzyme in bacteria. These three RNA polymerases are similar in structure and subunits but transcribe different types of genes[2]. RNA polymerases I and III are responsible for transcribing genes that code for ribosomal RNA, transfer RNA and various small RNAs[3,4]. RNA polymerase II (RNA Pol II) transcribes genes that code for proteins, producing messenger RNAs (mRNAs) and a large number of long non-coding (lnc) RNAs[1,5]. The transcription process that will be described below is the one catalyzed by this enzyme. The process of transcription of a gene by RNA Pol II is regulated on several levels starting with the initiation and transition into productive elongation, followed by pre-mRNA processing, e.g. splicing, until reaching the 3' end of a gene where cleavage, polyadenylation and RNA Pol II termination take place[1,6]. These processes will be described in more detail below.

1.1.1 Initiation and Elongation

Initiation of transcription is an important step in regulating the expression of a gene. At this step it is defined which genes are poised for transcription. The rate of RNA Pol II transitioning into productive elongation then define if the gene is expressed and at what rate[7]. In this stage, RNA polymerase binds to the gene promoter with the help of the general transcription factors (TF). These factors bind to the promoter, recognizing the TATA box, promote the opening of the double stranded DNA (dsDNA) and the passage of RNA polymerase to the elongation phase[8]. TF involved in RNA polymerase II transcription are called TFII. The association of transcription factors begins with the association of TFIID with TATA box, located 25 nucleotides upstream the transcription start site (TSS). This transcription factor causes physical distortion of DNA at this site, promoting the binding of other transcription factors and RNA Pol II, which constitute the transcription initiation complex[9]. Next, TFIIH, which contains a DNA helicase as one of its subunits, promotes unwinding of dsDNA and exposure of the template strand[6]. The opening of dsDNA leads to the formation of the transcription bubble and the synthesis of the transcript begins. For the shift of RNA Pol II to the elongation phase, C-terminal domain (CTD) phosphorylation of the largest subunit of RNA Pol II is required[10]. In humans, this domain consists of 52 tandem repeats of seven amino acids, including three serines which can be differentially phosphorylated. During the beginning of transcription, the serine in 5th position is phosphorylated by TFIIH, which has a kinase as one of its subunits[11]. From this moment on, the polymerase undergoes a series of conformational changes that increase its affinity to the DNA strand and allow it to be released from the transcription initiation complex[12]. After RNA polymerase transcribes a small RNA segment, it pauses downstream the TSS, due to the binding of DRB sensitivity- inducing factor (DSIF) and negative elongation factor (NELF)[13]. The entry of RNA Pol II into productive elongation is mediated by the cyclin-dependent kinase 9 (CDK9), a subunit of the positive transcription elongation factor b. The CDK9 phosphorylates NELF, which dissociates from the polymerase, DSIF and the serine in 2nd position of the CTD of the RNA Pol II, which restarts transcribing[14]. Enhancers also contribute for this process by recruiting cofactors that stimulate CDK9 or directly interfere with the pause and release of RNA Pol II, as bromodomain-containing protein 4 and p3000[15].

The initiation of transcription in eukaryotes is quite complex and involves several regulatory proteins, such as transcriptional activators or repressors which bind to the regulatory sequence of the gene[1]. The mediator complex is also needed to link RNA Pol II to regulatory proteins and transcription factors[16]. In

In order to initiation complex have access to the DNA, the action of chromatin modifying enzymes, i.e. chromatin remodeling complexes, as the chromatin structure remodeling complex[17], and histone modifying enzymes, such as histone acetyltransferase Gcn5[18], is required. There is not a single path for the association between the protein subunits, as it varies according to the gene to be transcribed[6,15].

During elongation, RNA Pol II moves irregularly, changing its speed and even pausing in some regions. This phase is controlled by elongation factors that bind to RNA Pol II soon after initiation, preventing it from dissociating from DNA before reaching the end of the gene. Chromatin remodeling complexes act on the chromatin structure to facilitate the passage of RNA Pol II and may be associated with this enzyme or recruited by other factors of the transcription complex to help any that is trapped[19].

The movement of RNA Pol II generates positive superhelical tension of the DNA ahead which hinders the opening of the DNA helix and facilitates the release of histones from DNA. In eukaryotes, topoisomerases help to release the positive superhelical tension. To compensate, behind the polymerase a negative superhelical tension is created and a DNA double helix with this conformation is more relaxed, allowing for instance the invasion of RNA molecules to hybridize with the antiparallel strand of homologous sequences to form so called R-loops[20], as will be discussed later.

1.1.2 Splicing

As most human genes contain non-coding intron sequences, their removal from the transcript synthesized is required. Splicing is the removal of non-coding regions (introns) so that the mRNA contains only coding regions, in a given open reading frame, for a given protein. The exons of a gene are interspersed with introns, which are often much longer regions in most of eukaryotes[21]. Thus, only a small part of the gene corresponds to the coding sequence for the protein. Each splicing event removes an intron in the form of a lariat and joins two exons through two sequential transesterification reactions. It is also this process that allows eukaryotes to synthesize more than one protein from the same gene, by allowing alternative splicing of intron/exon combinations in a regulated manner. Predominantly, splicing occur co-transcriptional but occasionally it happens after transcription[22,23].

The spliceosome is a complex formed by RNA molecules and proteins responsible for the splicing process. These RNA molecules are small nuclear RNAs (snRNAs) that recognize splice sites and are involved in the reaction. The snRNAs that constitute the spliceosome are U1, U2, U4, U5 and U6. Each is associated with at least seven proteins forming a small nuclear ribonucleoprotein (snRNP). For splicing to occur, the spliceosome must recognize the 5’ splice site, the 3’ splice site, and the branch point within the intron. Each splice site has a consensus sequence, which allows recognition of where splicing is to occur. U2 and U6 link 5’ splicing site and branch point, catalyzing the first transesterification. Due to the action of U5, the 5’ and 3’ splice junctions are approximated for the second transesterification to take place. After the reactions, snRNPs remain associated with the lariat, requiring rearrangements in RNA:RNA interactions and ATP hydrolysis to be released and used in a new splicing process. The exon junction complex then completes the splicing by linking the two exons[24–29].

1.1.3 3’ end processing and RNA Pol II termination

When the transcribing RNA Pol II progresses towards the 3’ end of a gene, it will reach the so called 3’ untranslated region (3’ UTR) in which, in most human genes, a sequence motif is found that has the consensus sequence AATAAA[30–32]. This signal is recognized in the transcript by RNA binding proteins and RNA processing enzymes, signaling pre-mRNA cleavage and polyadenylation as well as transcription termination. Although, these two mechanisms are not always interconnected. The cleavage stimulation factor (CstF) and cleavage and polyadenylation specificity factor (CPSF) are two very important protein complexes for the 3’ end processing. CPSF is recruited for the elongation complex,

interacting with the RNA Pol II body, while CstF binds to the CTD. When the AATAAA sequence is transcribed, CPSF binds to it, inducing polymerase pausing. CPSF also binds to CstF when it recognizes the GU-rich region downstream the cleavage and polyadenylation signal (CPAS). The pausing effect on the polymerase can also be caused by the chromatin structure or by hybridization of the nascent transcript to its DNA template[30,32–34]. Then, CPSF73, one of the components of the CPSF, cleaves the pre-mRNA and additional cleavage factors are recruited for the release of the transcript from the polymerase. The poly-A polymerase (PAP) is also recruited to add about 200 adenine nucleotides to the 3' end of the cleaved transcript, without the need for a template, forming the poly-A tail. As the poly-A tail is created, poly-A-binding proteins associate with it, determining its final size and protecting the mRNA from being degraded. Some remain bound to mRNA even after export to the cytoplasm, helping to address it to the ribosome[30,35].

Even though the mRNA has already been cleaved and polyadenylated, the RNA Pol II continues transcribing a few hundred more bases beyond the CPAS, eventually releasing from the template, the final step of transcription termination. There are two models for RNA Pol II transcription termination: the allosteric model and the torpedo model (Fig. 1.1). The allosteric model argues that transcription of the CPAS leads to polymerase conformational changes, with dissociation of elongation factors and association of termination factors. These alterations reduce RNA polymerase's processivity making it more prone to dissociate from its DNA substrate. According to the allosteric model, termination occurs through the limitation of the elongation efficiency of RNA Pol II, increasing the likelihood that the polymerase will dissociate. Previous studies have shown that the protein Pcf11, probably recruited by the polyadenylation signal in the nascent transcript, associates with the CTD and causes RNA Pol II to dissociate from DNA, proving that the poly(A) signal is sufficient to induce elongation complex disassembly independent of transcript cleavage[36–38]. In the torpedo model, exoribonuclease 2 (XRN2) is recruited by the CTD associated with 3' end processing factors. After the cleavage of the pre-CPAS transcript, this enzyme degrades the unprotected 5' end of the transcript produced by the polymerase after the CPAS. The collision of XRN2 with the polymerase leads to its dissociation from the template DNA strand, favoring the termination of transcription[31,34,39]. The arrest of RNA Pol II appears to play an important role in transcription termination in mammals, although it is not yet known if this is a general requirement. In addition to the association of CPSF with the polymerase, the pause of RNA Pol II is favored by the formation of hybrid RNA:DNA structures (R-loops), resulting from the invasion of the RNA transcript into the DNA and its hybridization with the template strand[33].

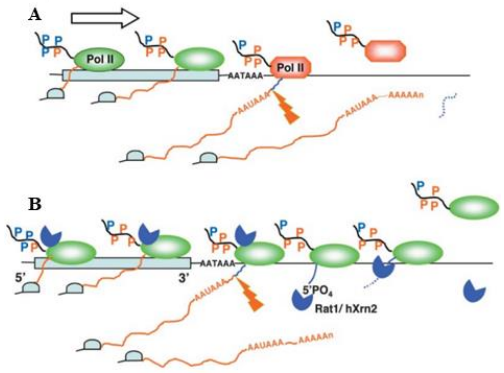


Figure 1.1 Two models of transcription termination by RNA Pol II. **A** Allosteric model. This model proposes a conformational change in the polymerase after detecting the polyadenylation site, weakening the ternary complex and favoring termination. **B** Torpedo model. In this case, the cleavage of the pre-mRNA forms a 5' phosphate end, providing an entry to XRN2. The exoribonuclease degrades the nascent transcript and, when it reaches polymerase, helps to displace it from the DNA template. Adapted from Luo *et al.* (2004)[39].

1.2 R-loops

As briefly introduced before, R-loops are nucleic acid structures that are formed by an RNA invading and hybridizing to one strand of a dsDNA, resulting in a displaced free single stranded DNA (ssDNA)[40,41]. These structures were first described *in vivo* by Drolet *et al.* in 1995 as a consequence of transcription in *Escherichia coli*[42]. Since then, it has been realized that R-loops can be found in higher bacterial, yeast and eukaryotic genomes. They are intermediates in several physiological processes, such as the initiation of mitochondrial DNA replication or the recombination of immunoglobulin class changes, although they also contribute to DNA damage and genomic instability[40,43].

R-loops distinguish from RNA:DNA hybrids that form transiently within the active center of the RNA Pol II by their size (between 100 and 2000 base pairs) and their higher stability[40]. The most accepted model for the formation of R-loops is the "thread-back" model. This model posits that R-loops result from dsDNA invasion by an RNA molecule produced by RNA Pol II during transcription, and they are not an extension of the transcription bubble[43]. The "thread-back" model is supported by the crystallography-resolved RNA Pol II structure, which shows the output of the transcript RNA molecule and the DNA template strand to exit via independent channels[44]. R-loop formation is favored by several factors such as high G content in the non-template sequence, negative supercoiling, and DNA strand nicks[40,41]. In addition, regions with a high GC skew have a high probability of R-loop formation, as there is a strong Watson-Crick base pairing between the RNA and the template chain, and the possibility of G-quadraplex formation of the displaced ssDNA chain[45]. The negative supercoiling found behind the RNA Pol II during transcription leads to an increased extension of RNA:DNA hybrids, because it weakens the dsDNA, leading to the separation of the DNA strands. Thus, both negative supercoiling and high G content promote the opening of dsDNA for RNA molecules to invade [41]. R-loops are thermodynamically more stable than dsDNA because they adopt a conformation between dsDNA (form B) and dsRNA (form A), as demonstrated by nuclear magnetic resonance and X-ray diffraction[46].

R-loops can be detected either directly or indirectly. Direct methods include electron microscopy, nucleic acid isolation and analysis of RNaseA resistance and RNaseH susceptibility, immunoprecipitation (IP) and immunofluorescence using the monoclonal antibody produced by the S9.6 hybridoma cell line. Indirectly, R-loops can be identified by mutation profiles caused by sodium bisulfite deamination or by human activation-induced cytidine deaminase (AICDA) in the misplaced ssDNA[40,47].

1.2.1 Factors that maintain R-loop homeostasis

The R-loops are involved in the regulation of cellular processes, as transcription or class switching but they may be involved in DNA damage also[48]. To maintain R-loops homeostasis, cells have evolved mechanisms that resolve or prevent their formation. One example is RNaseH enzymes, which specifically degrade the RNA moiety, independently of the sequence. There are two types of RNaseH enzymes, differing in their structure and specialized roles. RNaseH1 is a single polypeptide in which the N-terminal domain (hybrid binding domain) recognizes the RNA and the C-terminal domain contains the catalytic site of the enzyme. This enzyme is mainly responsible for degrading co-transcriptional R-loops. On the contrary, RNaseH2 is a multimeric protein, constituted by 3 subunits, in eukaryotes. RNaseH2a is the subunit that cleave the RNA strand of the RNA:DNA hybrid, while RNaseH2b and RNaseH2c are non-catalytic subunits, but their exact functions remain unclear[41,43]. RNaseH2b mediates the interactions between the proliferating cell nuclear antigen (PCNA), a protein involved in DNA replication and repair, suggesting a role of RNaseH2 in these mechanisms[49]. Both types of RNAseH can remove RNA primers during replication[41].

Another type of enzymes involved in R-loop resolution are RNA:DNA helicases such as Sen1 (yeast) or senataxin (SETX), aquarius (AQR) and DHX9 (human). The function of these enzymes is to unwind the RNA:DNA hybrids without destroying the mRNAs, playing a very important role during transcription. For example, SETX is responsible for unwinding the R-loops formed downstream of the poly(A) site, allowing the degradation of the post-CPAS nascent transcript by XRN2 and consequently mediate the efficient release of RNA Pol II, according to the torpedo model[41,50]. Indeed, it has been shown that SETX depletion leads to the accumulation of R-loops in the terminal region of the gene and negatively affects transcription termination[51].

Topoisomerases are also involved in controlling the formation of R-loops by reversing negative supercoiling of the DNA, however their action depends on the context. Depletion of topoisomerase I for instance can also lead to the accumulation of R-loops in long genes with a high transcription rate[52].

Another strategy to control the R-loops formation is to suppress proteins that promote their formation, as AtNDX in Arabidopsis[53] or Rad51 in eukaryotes. The study of the latter has special relevance because it showed that R-loops could also be formed in *trans*, with RNA molecules invading the dsDNA in a genomic region distant from the one where it was transcribed[54,55]. *Trans* R-loops formation is favored by an AGGAG repeat in the DNA, when it is negatively supercoiled[56]. *Trans*-induced R-loops pose a greater threat to genomic stability because transcripts from repetitive regions can form R-loops at various locations in the genome, promoting DNA damage[53].

Proteins that have no direct action on the formation or depletion of R-loops can also prevent their formation indirectly. RNA-binding proteins cover the nascent mRNA and simultaneously prevent it from hybridizing with the DNA strand being transcribed. Similarly, factors involved in RNA processing and biogenesis may contribute to the prevention of R-loop accumulation and the reduction of genomic instability[43].

1.2.2 R-loops in transcription regulation

R-loops play an important role in regulating some biological processes, such as transcription. Genome wide studies have shown the role of R-loops in transcription by identifying regions of R-loop formation. These regions correspond often to the promoters and termination regions of various human genes[43].

<u>Transcription Initiation</u>

In genes transcribed by RNA Pol II, there is evidence for an abundant formation of R-loops in CG-dinucleotide rich sequences (CpG islands), the majority present at 5' end of genes where they function

as promoter elements, some containing transcription start sites[57–59]. The CpG islands are characterized by the frequent absence of DNA methylation, however they can be methylated when associated to imprinting and tissue specific gene expression[60,61]. For most of the unmethylated CpG islands promoters, Ginno *et al.* (2012) demonstrated that they have a significant GC skew, conferring the ability to form R-loops upon transcription. They also suggest that the R-loop formation at CpG islands can prevent methylation because methyltransferases are not able to bind them[62]. R-loops may also promote or suppress the binding of transcription factors; yet, it is not clear in which context this happens. Previous studies indicated that the binding of transcription factors to R-loops depends on the sequence[40].

Class switch recombination

The class switch recombination (CSR) for immunoglobulins (Igs) occurs in germinal center B cells, after cytokine stimulation, changing the type of Igs produced by B cells, by modification of the constant region of the heavy chain[63]. Upstream of the constant region of each isotype (except for IgD) there are conservative sequences called switch (S) regions, expressed under a promoter inducible by a specific cytokine, which dives the production of a lncRNA. This lncRNA promotes the formation of R-loops, resulting in non-template ssDNA[64]. The AICDA deaminates cytosines located in single strand, which are converted to uracils[65]. Then, the enzymes of the base excision repair and mismatch repair pathways detect the uracils and create DNA DSB, predominantly resolved by non-homologous end-joining. The S regions are not homologous to each other and there is not a consensus sequence at the junctions of DNA fragments, so, probably, the CSR is not targeting a specific sequence, but a common structure forming at these *loci*[63].

Transcription Termination

The termination of mRNA transcription can also be regulated by R-loops[51,66]. In several genes, its terminal region has a high GC skew after the CPAS, which promotes R-loop formation, similar to the mechanism that occurs in promoters[43]. The formation of R-loops facilitates RNA Pol II pausing after the CPAS and, consequently, its release. Thus, the presence of R-loops contributes to an efficient termination, avoiding read-through. However, to prevent their accumulation, these R-loops must be resolved by SETX, and the nascent RNA attached to RNA Pol II degraded by XRN2[32,43,45]. Moreover, genome wide studies have shown that G-quadruplex forming sequences are prevalent in 3' UTR of genes, in regions of high gene density, with another nearby gene downstream[67,68].

To reinforce RNA Pol II arrest, R-loops in the terminal region of some genes can trigger antisense transcription through the formation of dsRNA, which recruits the RNA interference (RNAi) machinery and establishes repressive chromatin marks, notably H3K9me2. It is not yet known in this case how R-loops relate to the repressive marks of heterochromatin, since genome wide studies associate these secondary structures with H3K36me3 and H3K79me2, transcription activation marks[40,52].

1.2.3 R-loops in DNA damage and genomic instability

As mentioned before, R-loops are important to regulate transcription and related processes, however they have also harmful consequences, as DNA damage[48,69,70].

During the R-loop formation, a ssDNA is exposed, which is more prone to DNA damage. This strand is more susceptible to the action of AICDA and ApolipoproteinB mRNA editing catalytic polypeptide-like family enzymes, which convert C to U, making the ssDNA a substrate for the base excision repair enzyme uracil DNA glycosylase. Consequently, abasic sites are generated which can lead to base substitutions or can result in DSBs, after DNA replication[45,71,72].

In the absence of some RNA processing factors, like AQR and SETX, or inhibition of topoisomerase I, R-loops accumulate and are processed into DNA DSB by xeroderma pigmentosum complementation

group F and group G proteins. These endonucleases are recruited by transcription-coupled nucleotide excision repair (TC-NER) cockayne syndrome group B, upon the stalling of RNA Pol II, which suggests that TC-NER factors, responsible for correcting nucleotide errors during transcription, can affect negatively genome stability[73].

In addition, R-loops formed during transcription may interfere with DNA replication, originating a transcription-replication conflict (TRC) and leading to replication fork blockage and DSBs. However, the structural intermediates formed during TRC resolution remain unclear[74].

On the contrary, DNA damage can also lead to the formation of R-loops. For example, if transcription encounters a lesion in the DNA, the RNA polymerase is forced to pause, and processing factors associated with it are displaced. This facilitates the invasion of the transcript into the dsDNA, forming R-loops, which in turn trigger the DNA damage response[75].

1.2.4 R-loops in human diseases

Due to their ability to create genomic instability, the formation and non-resolution of R-loops is associated with some pathological conditions, namely neurodegenerative diseases and cancer. Regarding neurological diseases, R-loops form in genes with an abnormal expansion of repetitive DNA sequences, originating repeated expansion disorders[45]. Hexonucleotide repeat expanded in chromosome 9 open reading frame 72 (C9orf72 HRE), GGGGCC, is associated with a spectrum of neurological diseases such as amyotrophic lateral sclerosis and frontotemporal dementia[76,77]. The properties of C9orf72 HRE enable R-loop formation and ssDNA stabilization through G-quadruplexes[76]. Another example is Fridriech's ataxia which results in expansion of unstable GAA repeat in the first intron of the *FXN* gene, promoting the formation of R-loops, leading to RNA Pol II arrest and, consequently, decreasing the expression of this gene. Fragile X syndrome and fragile X-associated tremor/ataxia syndrome are also associated with the extension of CGG repeat in the 5' UTR region of the *FMR1* gene, decreasing its expression. Loss of protein function associated with R-loop resolution is also associated with neurodegenerative diseases. In case of mutations in SETX that negatively affect its function, neuronal differentiation is affected and ataxia with oculomotor apraxia Type 2 develops[43].

Cancer is a group of several complex diseases that are characterized by loss of tumor suppressor function, oncogene signaling, high levels of mutagenesis and DNA damage. As already mentioned, R-loops cause genomic instability, interfering with essential molecular processes such as replication and transcription[41]. In a cancer cell, genes with a higher transcription rate are more prone to DNA damage through the formation of R-loops and, depending on the genes in which they form, lead to a specific mutagenic phenotype. Mutations in cancer susceptibility factors, such as BRCA1 and BRCA2 in breast cancer, can also lead to the accumulation of R-loops[40].

1.3 Single-molecule imaging in study of transcription and R-loops

Structural, biophysical and biochemical methods describe how a set of molecules behave, estimating average values for the parameters studied in a population of cells. Most of the techniques used so far to study transcription dynamics and R-loop function belong to this group, such ChIP/DIP or native elongating transcript sequencing (NET-Seq)[78].

In this project we use single-molecule sensitive live cell imaging, which constitutes a different approach to the study of transcription dynamics and the influence of R-loops on it. Single-molecule imaging has a very high sensitivity and allows to observe single molecule tracking in live cells. One of the advantages of this technique is that the reaction which will be detected does not need to be synchronized, as it happens in multiple-molecule measurements. Besides, single-molecule measurements give information about the fluctuations and distributions of dynamic and kinetic parameters. Single-molecule analysis

also allows to monitor individually each input and output of single events of protein reactions, as the binding and dissociation of a ligand. Despite these advantages, this method requires the statistical analysis of many events, which is a laborious task[79].

One of the visualization strategies in single-molecule imaging is the use of fluorescent proteins to monitor the behavior of the molecule, useful to perform kinetical studies of transcription. The fluorescent protein is directly linked, through genetic engineering, to a particular protein, which recognizes a specific sequence in a mRNA molecule. The fluorescent protein used should be bright, have small fluctuations in signal emission intensity, be small and to not disturb the molecule under study[80]. In this project, a strategy similar to that already used for visualizing pre-mRNA splicing was used[81]. The HEK 293 cells are transfected with a reporter gene, in which there is a cassette coding for 24 tandem repeat stem-loops recognized by the coat protein of bacteriophage PP7 into exon II and a cassette coding for 25 stem-loops recognized by the 22 amino acids of the binding domain of bacteriophage λ antiterminator protein N (λN_{22}) after the CPAS. For mRNA molecules to be visualized, the cells were also transfected with PP7 protein fused with mCherry fluorescent protein and λN_{22} protein fused with green fluorescent protein (GFP). Thus, when the transcription of the reporter gene is induced in live cells, the PP7 and λN_{22} proteins recognize and bind to the stem-loops transcribed, allowing to follow the production and release/degradation of transcripts in real-time. By analyzing the variations of fluorescence, it is possible to infer kinetic parameters of the RNA Pol II, like its velocity during transcription elongation.

1.4 Objectives

The aim of the project is to investigate the effect of R-loop formation on the process of transcription and RNA processing. The plan was to provide a kinetic description of the pre-mRNA 3' end processing and transcription termination as well as to determine the effect of R-loops on the dynamics of these processes as an important regulatory mechanism. Furthermore, we assessed the influence of R-loop formation on the efficiency of splicing, when they localize in the 3' end of the intron. Finally, we inferred how the co-transcriptional processing of pre-mRNAs influences the dynamics of transcription termination.

To achieve the proposed goals, we used spinning disk confocal microscopy with single-molecule sensitivity to detect individual pre-mRNA transcripts of reporter genes equipped with arrays of repetitive stem-loop forming sequence elements. Upon transcription, these RNA stem-loops structures are recognized and bound by specific proteins derived from bacteriophage RNA binding proteins fused with green/red fluorescent proteins, allowing microscopic visualization of individual RNA transcripts. We generated reporter gene variants bearing an R-loop forming sequence (RFS) that was identified in the beta-actin gene post-CPAS or in the 3' end region of the intron[66] or a non-R-loops forming control sequence (NRFS) post-CPAS. These constructs allowed to investigate the effect of R-loop formation on the kinetics of pre-mRNA processing and transcription termination. We combined the reporter genes with an inducible RNAseH system to test the effect of immediate R-loop degradation on the dynamics of pre-mRNA 3' end processing and transcription termination.

2. Methods

2.1 Cell culture

The cells used in this project were Flip-In™ human embryonic kidney (HEK) 293 cell line (Thermo Fisher Scientific), which contain in their genome a single integrated Flipase Recombination Target (FRT) site. Using the Flip-In system 6 cell lines were generated with different reporter genes. All cell lines were grown as monolayers and were maintained in T25 flasks (Thermo Fisher Scientific), in Dulbecco's Modified Eagle Medium (DMEM) [Gibco by life technologies] supplemented with 10% (v/v) fetal bovine serum (Gibco by life technologies) and 1% (v/v) 200 mM L-glutamine (Gibco by life technologies). Cells were kept in a 5% CO2 humidified incubator at 37 ºC. To maintain the selective pressure, only allowing cells which carry the reporter gene to grow, Blasticidin 150 µg/ml and Hygromycin B 200 µg/ml (Invivogen) were added.

To perform the live cell experiments as well as RNA and protein extractions mentioned below, the cells were seeded and maintained in the same conditions as described but without selection antibiotics, to diminish the variability of the experiments. For live cell microscopy experiments, the cells were seeded on 35 mm petri dishes with 10 mm glass bottom (MatTek Corporation) and maintained in the same conditions but using DMEM without phenol red (Gibco by life technologies), to avoid interference with the fluorescence signal.

2.2 Generation of reporter gene cell lines and transfections

To generate the cell lines for this study, Flip-In™ HEK 293 cells were transfected with 0.1 µg of pcDNA5/FRT/TO expression vector containing the reporter gene as well as the FRT site and 0.9 µg of the Flp recombinase expression plasmid, pOG44. The Flp recombinase mediates the insertion of the reporter gene into the genome at the integrated FRT site through site-specific DNA recombination. Stable cell lines expressing the gene of interest were selected using Hygromycin B. Transcription of the reporter genes is driven by the human cytomegalovirus (CMV) promoter and regulated by two copies of the Tet operator, a system derived from the tetracycline-resistance operon (TetO). The Flip-In™ HEK 293 cells express already the Tet repressor, which binds to the Tet operator, inhibiting the expression of the reporter gene. So, for the reporter gene to be transcribed, it is necessary to add Doxycycline (Dox), which binds to the Tet repressor to prevent its binding to the TetO[82].

The reporter genes are derived from mouse IgM gene, exons M1 and M2 and the intron between, but the polypyrimidine (PY) tract from adenovirus major late or derived from the C4-M1 exon junction of IgM gene, with a degenerated (non-consensus) intron exon junction (PY1wsj). Intron was extended by 1.7 kbp from fragments derived from the first intron of the mouse RNA Pol II gene before and after the MS2 binding sites [81]. Also, the reporter gene encodes a cyan fluorescent protein (CFP) with peroxisomal targeting signal in exon II. In all the reporter genes, there are a cassette of 24 tandem repeated PP7 stem-loops in exon II, a cassette of 25 tandem repeated BoxB stem-loops after the CPAS and 24 tandem repeat MS2 stem-loops in the intron (Fig. 2.1). The stem-loops are specifically bound by the coat protein of bacteriophage PP7, the λN_{22} protein and the MS2 bacteriophage coat protein, respectively. Only in the cell line used for the calibration of λN_{22}-3xGFP signal labeled transcripts, the reporter gene had the 25xBoxB stem-loops in exon II (Annex 1). A fluorescent protein was fused in-frame with the C-terminus of each binding protein, mCherry9ikkkk with MS2 protein and GFP with λN_{22} protein, so that the transcription before and after the CPAS could be detected by microscopy [83]. Each fused protein contains also a nuclear localization signal that confines them to the nucleus. Thus, when the transcription of the reporter gene is not induced, diffused fluorescence is detected in the nucleus. Upon induction of the transcription with Dox, a fluorescent dot is detected, corresponding to the transcription site (TS).

Plasmids were transfected according to the supplier's protocol with Lipofectamine 3000 reagent (Invitrogen). Expression of the reporter genes was induced with 0.01 up to around 0.08 µg/ml of Dox (Sigma, St Louis, MI). The RNaseH1 used is bacterial and coupled to the ligand binding and nuclear translocation domains of a glucocorticoid receptor (GR) and to near-infrared fluorescent protein (iRFP) [RNaseH1-GR-iRFP]. When the GR is not associated to any ligand, the construct is cytoplasmatic and the GR is sequestered in a multimeric chaperone complex. Upon ligand binding, the GR dissociates from the chaperone complex and the construct migrates into the nucleus, where RNaseH degrades the R-loops. Thus, it is possible to control its migration to the nucleus by adding 10^{-7} mM triamcinolone acetonide (TA), which binds to the receptor.

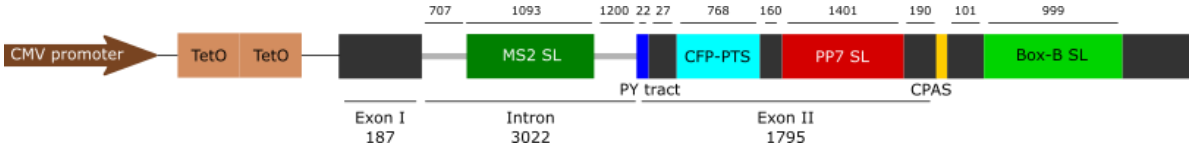


Figure 2.1 Scheme of the reporter genes construction. The lengths of each region are in base pairs (bp). The MS2 stem-loops (SL), PP7 stem-loops (SL) and BoxB stem-loops (SL) are inserted in the intron, exon II and post-CPAS, respectively. The reporter gene also encodes a cyan fluorescent protein with peroxisomal targeting signal (CFP-PTS) in exon II.

2.3 Live cell spinning-disk confocal imaging and image analysis

Live cell imaging experiments were performed on a 3i Marianas SDC spinning disk confocal imaging system (Intelligent Imaging Innovations, Inc.), using a similar microscopy setup previously described[84]. The system is based on an Axio Observer Z1 inverted microscope (Carl Zeiss MicroImaging, Inc.) equipped with a Yokogawa CSU-X1 spinning-disk confocal head (Yokogawa 30 Electric, Tokyo, Japan) and 100 mW solid state lasers (Coherent, Inc.; Santa Clara, CA) coupled to an acoustic-optical tunable filter. The axial position of the sample is controlled with a piezo-driven stage (Applied Scientific Instrumentation, Eugene, OR). Each MatTek dish was placed in an incubation chamber (Pecon P-Set 2000; Pecon GmbH, Erbach, Germany) mounted on the microscope stage and connected to CO_2 (Pecon, CO_2 module S) and humidity (Pecon, Heating Device Humidity 2000) controllers. The whole microscope body except lasers, camera and spinning disk head are maintained inside a large plexi glass environmental chamber (Pecon, Erbach, Germany). The temperature in both the microscope and top stage incubation chambers is controlled by a common unit and set to 37ºC. The environment inside the top stage incubation chamber is further set to 5% CO_2 and 100% humidity. Samples were illuminated with λ = 488 nm for GFP, λ = 561 nm for mCherry and λ = 640 nm for iRFP. Images were acquired using a 100x 10 (Plan-Apo, 1.4 NA) oil immersion objectives (Carl Zeiss, Inc.) under control of Slidebook 6.0 software (Intelligent Imaging Innovations, Denver, CO). Three-dimensional (3D) time-lapse image stacks of 8 optical slices separated by 0.32 µm were collected every 4 s for 5 min, with exposure acquisition times of 50 ms. Digital images (16-bit) were acquired using a back thinned air-cooled electron-multiplying charge-coupled device (EMCCD) camera 15 (Evolve 512, Photometrics, Tucson, AZ). To detect the transcription of individual RNA, 0.01-0.08 µg/ml Dox was added, and cells were incubated during 1 h at 37 ºC, 5% CO_2. For calibration experiments, the transcription was induced with 3 µg/ml Dox to increase the rate of RNA Pol II initiation/elongation from the promoter, thus, more transcripts are synthetized and can be detected disperse in the nucleus. The cells were incubated in the same conditions.

For the analysis of 3D time-lapse sequence, a software for spot tracking and quantification was used, STaQTool[85]. This software was developed to automatically track the TS in the cell nucleus over time and measure its total fluorescence intensity (TFI) by Gaussian fitting at the Z plane corresponding to the highest intensity value. Then is generated a plot of TFI over time, representing the formation of transcripts from the reporter gene. The coordinates and TFI values are saved in an Excel file. To know

which TFI values correspond to a cycle of transcription of the stem-loops calibration experiments were performed (Annex 1). In these experiments, once the amount of Dox added was much higher (3 μg/ml), the rate of transcription of the reporter gene was higher too, so that it was possible to detect multiple released labeled transcripts in the nucleus. The cells were imaged under a spinning disk microscope in 2D, every 0.5 sec for 120 timepoints. We plot a histogram of the TFIs measured for 549-553 of these transcripts, then fit a Gaussian over the histogram and determine the peak Xc value of the Gauss curve. The TFI values on the X-axis corresponding to the Xc value is the mean TFI of labeled transcripts and this is the value we take for calibration. Then, all the TFI values measured in the live cell 3D time series were converted to number of transcripts. The TFI values which correspond to a cycle of transcription of the stem-loops were selected and plot against time. To calculate RNA Pol II speed, we measure the time that the TFI takes to increase between background levels of fluorescence (beginning of the transcription of stem-loops sequence) and a plateau, which corresponds to the transcription of a complete array of stem-loops. Knowing the length of this sequence (in nucleotides) and the time it takes to be transcribed, it is possible to calculate the speed of RNA Pol II in that region of the reporter gene. The time between the beginning of the transcription of stem-loops sequence and the cleavage, when the fluorescence signal decreases, was also measured. This was calculated for the PP7 stem-loop sequence and Box B, when its signal was detected (Annex 2). To compare the time measured for the transcription of the PP7 stem-loops between the reporter genes IgM-1.7k-PY, IgM-1.7k-PYpA-baRFS and IgM-1.7k-PYpA-NRFS we used a paired t-test.

2.4 RNA extraction and qPCR

To quantify the ratios of spliced and unspliced transcripts produced from the reporter genes, total RNA isolation was performed using TRIzol reagent (Life Technologies), according to manufacturer's protocol. The transcription of the reporter genes was induced with 5 μg/ml of Dox and cells were incubated for 2h, at 37 ºC, 5% CO_2.The synthesis of cDNA was performed using cDNA synthesis kit (NZYTech), with oligo-dT and random primers.

The cDNA was then used as a template in qPCR reactions, at a 1:15 dilution. The qPCR was performed in ViiA Real Time PCR machine (Applied Biosystems, CA, USA), using Power SYBR Green Master Mix (Applied Biosystems, CA, USA). The relative RNA expression we estimated as $2^{(\text{Ct reference} - \text{Ct sample})}$, where Ct reference and Ct sample are mean threshold cycles of RT-qPCRs done in duplicate of the PCNA housekeeping gene and the gene of interest (sample). We also calculated the percent spliced in (PSI), which indicates the efficiency of splicing a specific exon into the transcript population of a gene[86]. All primer sequences are presented in Annex 3.

2.5 RNA interference

To knockdown XRN2, by RNAi, a short hairpin RNA (shRNA) was used. HEK 293 cells harboring the IgM-1.7k-PY reporter gene were seeded on T25 flasks and transfected with of 3 μg of plasmid pSUPER-puro-shXrn2 (OligoEngine). As controls, three T25 flasks were prepared with the same cell line: two not transfected and one transfected with the empty vector. The shRNA transfected cells and controls were selected with 0.5 μg/ml of puromycin for 48h. Then, cells were split into 35 mm petri dishes for microscopy experiments and a six well plate (Thermo Fisher Scientific) for western blot.

2.6 Western blot

Western blot was performed to confirm the knock-down of XRN2 by RNAi in HEK 293 cells harboring the IgM-1.7k-PY reporter gene. Whole cell protein extracts were prepared by cell lysis with sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) buffer (80 mM Tris-HCL pH 6.8, 16% glycerol, 4.5% SDS, 450 mM DTT, 0.01% bromophenol blue) with 200 U/ml benzonase (Sigma) and 50 μM MgCl2 and boiling for 5 min. Equal amounts of protein extracts were resolved by SDS-PAGE

and transferred to a nitrocellulose membrane. Through dry transfer, with an iBlot System (Invitrogen). The primary antibodies used were anti-XRN2 (1:5000, Bethyl Laboratories). Alpha tubulin was used as a loading control. The horseradish peroxidase (HRP) –coupled secondary antibody used were goat anti-Mouse-HRP conjugate (1:5000, Biorad) and goat anti-Rabbit-HRP conjugate (1:5000, Biorad). Protein detection was achieved using enhanced chemiluminescence substrates, either less sensitive (GE Healthcare Life Sciences) or more sensitive (Thermo Fisher Scientific) depending on the abundancy of the target proteins.

3. Results

The RNA Pol II is an essential enzyme responsible for transcribing the genetic information encoded in DNA into RNA, with high efficiency, in eukaryotic cells. During transcription, the kinetics of RNA polymerase is affected by the characteristics of the sequence, epigenetic modifications and association with transcription factors[87]. R-loops are structures that influence transcription, however, how they affect transcription dynamics and RNA Pol II kinetics is not well understood. With single-molecule sensitive imaging of individual transcripts, over time in 3D, it is possible to determine with high temporal resolution several kinetic parameters, such as the RNA residence time, the RNA lifetime and the transcription rate of RNA Pol II. Single molecule sensitive imaging data are complementary to data obtained with biochemical methods, which demonstrate where are they formed and what are their characteristics[51,67,68].

To analyze the kinetic behavior of RNA Pol II in transcription termination, a reporter gene was engineered, named IgM-1.7k-PY (Fig. 3.1A), which was integrated into a HEK 293 cell line via the Flp recombinase system. Expression of this reporter gene is controlled by a CMV promoter and a Tet-on system[82], using a concentration of Dox between 0.01-0.08 μg/ml. For the microscopic visualization of transcription in live cells we used MS2 like RNA stem-loop systems. These systems are composed of tandem arrays of up to 25 small RNA stem-loop sequences. Each stem-loop is bound by a specific protein which can be tagged with a fluorescent protein. The reporter genes used in this study are composed of two exons separated by an intron from the mouse IgM gene. Into the intron there are 24 tandem repeat MS2 stem-loops which are recognized by the MS2 bacteriophage coat protein. Exon two encodes 24 PP7 tandem repeat stem-loops bound by the coat protein of bacteriophage PP7. After the CPAS, 25 BoxB tandem repeated stem-loops were inserted, which are bound by the λN_{22} protein. Upon induction of the transcription with Dox, the specific proteins bind to the stem-loops present in the transcript and a fluorescent dot is detected, corresponding to the TS.

3.1 Determining RNA polymerase II dynamics in transcription termination on a reporter gene

First, the transcription speed of RNA Pol II at the different stem-loop arrays in the reporter gene was determined. As briefly described before, the reporter gene contains PP7 stem-loops in exon II and Box B stem-loops downstream the CPAS. We transfected the cell line HEK 293 with mCherry-PP7 and λN_{22}-GFP expression plasmids to fluorescently label the PP7 and BoxB stem-loops. The cells were imaged under a spinning disk microscope over time in 3D. A z-stack of 8 slices was acquired every 4 sec for 75 timepoints. As shown in the Annex 2, the fluorescence signal corresponds to the diffuse unbound fluorescently tagged mCherry-PP7 and λN_{22}-GFP proteins which accumulate in the nucleus because of the presence of a nuclear localization sequence. The intense fluorescent dotlike signal for the mCherry-PP7 protein is the TS. The TS position is stable over time, and just the intensity fluctuates, indicating de-novo transcription and finished mRNAs leaving the TS. During imaging, it was seen an increase in the fluorescence signal at the TS that corresponds to the binding of mCherry-PP7 or λN_{22}-GFP proteins to the target sequence on nascent single RNA transcripts, as soon as they emerged from RNA Pol II. Then, the fluorescence signal stabilizes, when the stem-loops sequences are completely transcribed, and decreases to the background levels of fluorescence, corresponding to the release/cleavage of the RNA. Using the software STaQTool (version 1.1, 2016)[85], the 3D position of the TS is determined for each timepoint, the fluorescence dotlike signal is fitted in 3D with a Gaussian function, according to its size, and the TFI is automatically calculated. Next, the software plots the TFI against the timepoints, constructing a graph which shows the fluorescence fluctuations. These cycles correspond to the increase of the fluorescence signal from the background level, stabilization and decrease again to the background level of fluorescence, reflecting what is observed by microscopy.

To determine if this variation in fluorescence intensity corresponds only to the synthesis of a single transcript, calibration experiments were performed (Annex 1). They consist in transfecting the HEK 293 cells only with mCherry-PP7 or λN_{22}-GFP and induce the transcription with a higher concentration of Dox (3 µg/ml). Since, the BoxB stem-loops after the CPAS are not always transcribed, and it was not detected the release of transcripts in the nucleoplasm, as we show later, a reporter gene with the BoxB stem-loops sequence inserted in the exon II was used for calibration of the λN_{22}-GFP signal. By live cell microscopy, we acquired only one 2D image every 0.5 s for 1 min. Several free transcripts are visualized in the cell nucleus, identified by a fluorescence dotlike signal that is not positionally stable over time. STaQTool detects each transcript in each timelapse and calculates its TFI. The Gaussian fit analysis of all the TFI values measured for one type of RNA label, allows us to determine the average TFI value for a single transcript labeled by mCherry-PP7 or λN_{22}-GFP. Using the mean TFI value, the TFI values obtained for the transcription cycles in 3D timeseries can be converted to number of transcripts. For the calculation of RNA Pol II speed, it is necessary to measure the elapsed time since the beginning of the increase of the fluorescence signal until the plateau for each transcription cycle, corresponding to the time that RNA Pol II takes to transcribe the stem-loops sequence.

After inducing the transcription of the reporter gene represented in Fig. 3A, by live cell microscopy, we observed only the mCherry-PP7 signal at the TS, and no λN_{22}-GFP signal (Fig. 3.1B and C). This means that RNA Pol II is producing only a pre-CPAS transcript, or, if a transcript is being synthetized after the CPAS, it is immediately degraded. Assuming the torpedo model, immediate XRN2 degradation of post-CPAS transcripts makes it impossible for the λN_{22}-GFP to bind to the stem-loops present. As explained above, the images were processed in the STaQTool software, to determine the position and the TFI of the TS fluorescence signal over time and, the cycles of fluorescence intensity variation corresponding to the synthesis of a single transcript were selected (n=22). On average, the time measured for the increase of the fluorescence intensity between the background levels and a plateau, that is the time for the transcription of PP7 stem-loops sequence, was 17 ± 6 s. We also measured the time between the beginning of the transcription of the PP7 stem-loops and the cleavage (end of the plateau), obtaining 35 ± 10 s. Considering the 1.401 kb length of the 24xPP7 stem-loops, we calculated a transcription rate of 4.94 kb/min for the RNA Pol II. This result is in very good agreement with previous studies, showing that the normal RNA Pol II speed during elongation is between 2.00 and 6.00 kb/min[87–89].

To perform these measurements it is necessary to use low concentrations of Dox (0.01 up to around 0.08 µg/ml), to achieve a low RNA Pol II firing rate that allows the visualization of one transcript at a time, so the begining and the end of transcription of stem-loop sequences can be defined. Concentrations higher than 0.08 µg/ml increase the polymerase density in the reporter gene. In this case, transcription after the CPAS is detected through the binding of λN_{22}-GFP to the BoxB stem-loops (Fig. 3.1D). This indicates that a high polymerase density influences the site and timing of transcription termination, leading to a delay in RNA Pol II release from the template and transcription of the BoxB stem-loops sequence.

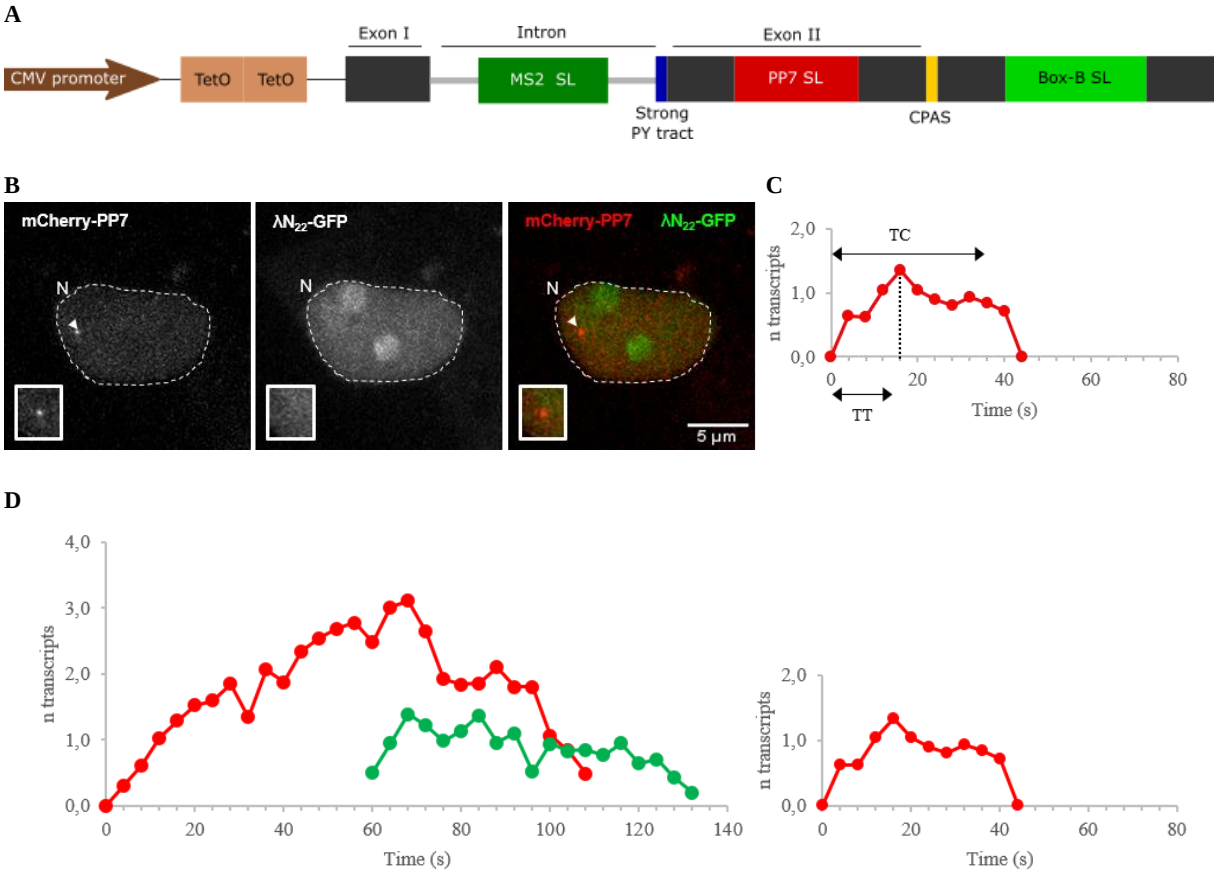


Figure 3.1 Transcription termination in IgM-1.7k-PY reporter gene**. A** Scheme of the IgM-1.7k-PY reporter gene under control of a CMV promoter and two Tet-operator sequences. Binding sites for mCherry-PP7 (PP7 SL) and λN_{22}-GFP (BoxB SL) were inserted in the exon II and after CPAS, respectively. Bindings sites for MS2 (MS2 SL) were inserted in the intron. The interaction of fluorescently tagged RNA stem-loop binding proteins with the respective stem-loops allows the visualization of the RNAs. **B** Representative image of a HEK 293 cell expressing transcripts tagged with PP7-binding sites in the exon II but not with λN_{22} binding sites. The nucleus (N) is delimited with a dashed line and the TS of the reporter gene is indicated in each cell with an arrowhead. Corresponding images depicting the diffraction limited objects in the highest intensity plane are shown below. **C** The number (n) of transcripts was plotted over time (s) in line graphs. This line plot depicts a complete cycle of fluorescence gain and loss present in a time-lapse series. The time of transcription (TT) and the time until cleavage (TC) are also indicated in the graph. **D** Line plots of number (n) of transcripts against time (s), comparing the synthesis of more than one transcript simultaneously (left) and the synthesis of a single transcript (right).

3.2 RNaseH decrease the efficiency of transcription termination

After measuring the RNA Pol II speed and detecting that, in an immediate transcription termination for our reporter gene, only the PP7 labelling sequence is transcribed, based in previous studies, we decided to investigate if there was a possibility of R-loop formation after CPAS of our reporter gene and what would be the effect of its removal. Skourti-Stathaki *et al.* (2011) showed, through ChIP, that RNA Pol II accumulates near the CPAS. They also observed, after knocking down SETX (RNA:DNA helicase involved in R-loop resolution) an additional increase in polymerases at this site. Using the S9.6 antibody, which recognizes RNA:DNA hybrids, the R-loop formation near the CPAS of the β-actin gene was detected. These data led them to conclude that the R-loop formation in the transcriptional pause regions is essential to pause RNA Pol II downstream the CPAS prior to termination, in the β-actin gene [51]. Therefore, we analyzed bovine growth hormone (BGH) CPAS sequence[90], present in our reporter gene, for its G content and GC skew. The 107 bp sequence has a positive GC skew of 0.475, representing the richness of G over C in the coding strand[91], and a GC content, percentage of C and G in a strand, of 57%, meaning that the sequence from the BGH CPAS is prone to R-loop formation (Fig. 3.2A).

Considering the characteristics of this sequence, we hypothesized that R-loops form in this region and contribute to the efficiency of RNA Pol II transcription termination. To test this, we reasoned that RNaseH transfection would degrade the R-loops formed in this region, thus resulting in RNA Pol II not stopping after transcribing the CPAS and in non-immediate transcription termination. This could be verified by detecting a readthrough, transcribing the BoxB stem-loops sequence. To test this hypothesis, we used the RNaseH1-GR-iRFP. The GR, in the absence of any ligand, resides in the cytoplasm where it is sequestered in a multimeric chaperone complex. The GR is dissociated from the chaperone complex, upon ligand binding, and migrates into the nucleus, where it interacts with specific DNA sequences in the regulatory regions of target genes and modulates their expression[92]. Thus, when we add TA, a synthetic corticosteroid, it binds to the GR binding domain of the RNaseH1-GR-iRFP construct, promoting the migration of the construct to the nucleus. The iRFP enables the visualization of the construct migration to the nucleus, which ensures that the experience ran successfully (Fig. 3.2B).

We repeated the experiment described above, transfecting the cells with the mCherry-PP7, λN_{22}-GFP and RNaseH1-GR-iRFP encoding plasmids. After inducing the expression of the reporter gene and adding TA, both signals of mCherry-PP7 and λN_{22}-GFP were observed at the TS in 27% of the single transcription events detected (n=22) [Fig. 3.2C-F]. This contrasts with the result detected before where all cells showed only the TS labeled with mCherry-PP7. These data indicate that, in the presence of RNaseH1, the transcription termination is not completely immediate, as verified by the detection of the transcription of BoxB stem-loops by λN_{22}-GFP. In addition, the time needed for RNA Pol II to transcribe the PP7 and BoxB stem-loops and the time until the cleavage of the transcripts were measured. When only PP7-binding sites are transcribed in cells transfected with RNaseH1-GR-iRFP, RNA Pol II takes 20 ± 6 s to transcribe PP7 stem-loops, at a transcription rate of 4.23 kbp/min, similar to the result obtained previously. The pre-CPAS transcript is cleaved after 38 ± 11 s after RNA Pol II started transcribing the PP7 stem-loops. However, when BoxB stem-loops are transcribed in cells overexpressing RNaseH1, RNA Pol II takes 12 ± 6 s to transcribe PP7 loops, corresponding to a transcription rate of 6.82 kbp/min. Thus, in 27% of the cells overexpressing RNaseH1, a higher RNA Pol II speed during transcription of PP7 stem-loops leads also to the transcription of the BoxB stem-loops. In this situation, the cleavage of the pre-CPAS transcript happens 36 ± 4 s after the beginning of the transcription of PP7 stem-loops and the BoxB stem-loops (0.999 kbp) are transcribed in 10 ± 6 s, at a transcription rate of 5.83 kbp/min.

The alteration of RNA Pol II speed in the PP7 stem-loop sequence in 27% of cells transfected with RNaseH1-GR-iRFP suggests that RNaseH1 acted on the R-loops formed before the PP7 stem-loops. In fact, despite high GC content and GC skew contribute to R-loop formation, these are not obligatory factors. Therefore, since R-loops may lead to the slowdown of RNA Pol II[93] and if the action of RNaseH1 has an effect on its speed during PP7 stem-loops transcription, then, it may act on R-loops formed upstream of this sequence.

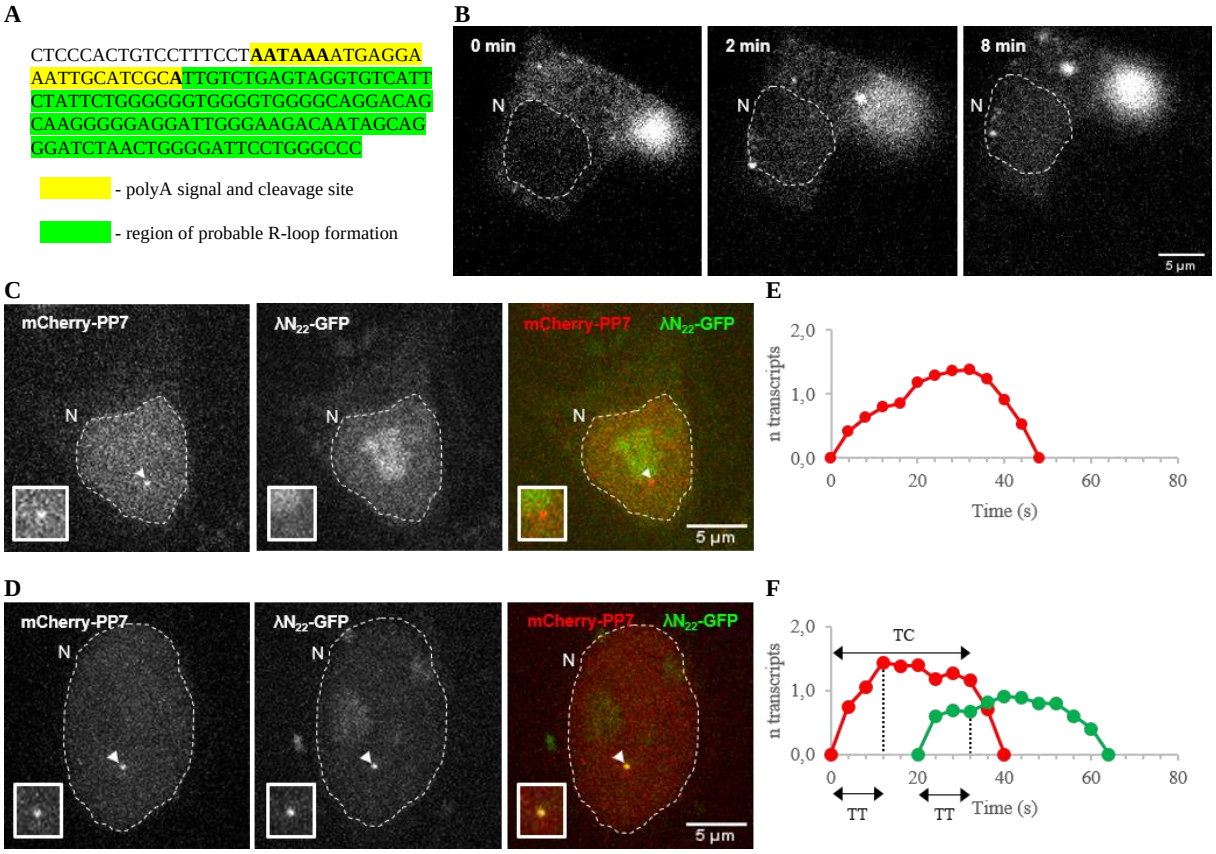


Figure 3.2 Influence of RNAseH1 on the reporter gene transcription. **A** Sequence around the CPAS from the BGH gene. The CPAS is highlighted in yellow and the sequence between the CPAS and the BoxB stem-loops is highlighted in green. The poly(A) signal is in bold. **B** Microscopy images showing the migration of RNaseH1-GR-iRFP to the nucleus, 0 min, 2 min and 8 min after TA treatment, detected by the migration of the iRFP fluorescence signal from the cytoplasm to the nucleus. The nucleus (N) is delimited by a dashed line. **C** and **D** Representative images of HEK 293 cells transcribing the IgM-1.7k-PY reporter gene under RNaseH1overexpression. The nucleus (N) is delimited with a dashed line and the TS of the reporter gene is indicated in each cell with an arrowhead. Corresponding images depicting the diffraction limited objects in the highest intensity plane of acquired z-stacks are shown in the insets. **C** In 73% of the cases, the TS is only labeled by mCherry-PP7 exon label. **D** The TS is labeled with both mCherry-PP7exon and λN_{22}-GFP post-CPAS labels and represents the 27% of cases where both labels were detected. **E** and **F** Representative single transcript fluorescent intensity graphs in cells expressing the IgM-1.7k-PY reporter gene with RNaseH overexpression. The calibrated number of transcripts (n) was plotted over time (s). **E** A cycle of a transcription representing 73% of the cases with the TS labeled only with mCherry-PP7. **F** A cycle of a transcript representing the 27% of cells with the TS-labeled with mCherry-PP7 and λN_{22}-GFP.

3.3 Formation of R-loops post-CPAS does not affect transcription termination

The BGH-post-CPAS sequence has never been directly shown to form R-loops, therefore we sought to introduce a sequence element from the β-actin gene that has been shown to induce R-loop formation, downstream the CPAS[94] (Fig. 3.3A). Following this, we used the same live cell imaging setup and tested if transcription termination is similarly efficient in the new reporter gene, IgM-1.7k-PYpA-baRFS. We observed the transcription of the sequence upstream the CPAS, labeled with mCherry-PP7 in all the cells imaged (n=22) but we did not detect a fluorescent signal for the labeling of the post-CPAS RNA (Fig. 3.3B and C). So, the transcription termination seemed to be immediate and the results obtained are similar to the ones for the IgM-1.7k-PY, with the BGH-CPAS site. Kinetic parameters were also determined, as already described. The average time measured for the transcription of the PP7 stem-loops in the IgM-1.7k-PY-RFS-postCPAS is 14 ± 6 s with a transcription rate of 6.04 kbp/min, identical to the value calculated when the transcription termination is immediate. So, when an RFS is inserted after the CPAS, the transcription is immediate. The cleavage is also efficient because it happens 33 ± 9 s after RNA Pol II started transcribing the PP7 stem-loops sequence.

We also found a similar behavior for IgM-1.7k-PY and IgM-1.7k-PYpA-baRFS, when their expression was induced with a Dox concentration of 0.08 µg/ml. Here again, transcription after the CPAS in the new reporter gene was detected through the binding of BoxB stem-loops by λN_{22}-GFP (Fig. 3.3D), showing a readthrough of RNA Pol II post the CPAS sequence probably due to an increase in the polymerase density at the 3' end of the reporter gene. This result also indicates a functional similarity between CPAS region of the β-actin gene and the BGH-pA region, given similar behavior in the presence of higher concentrations of Dox.

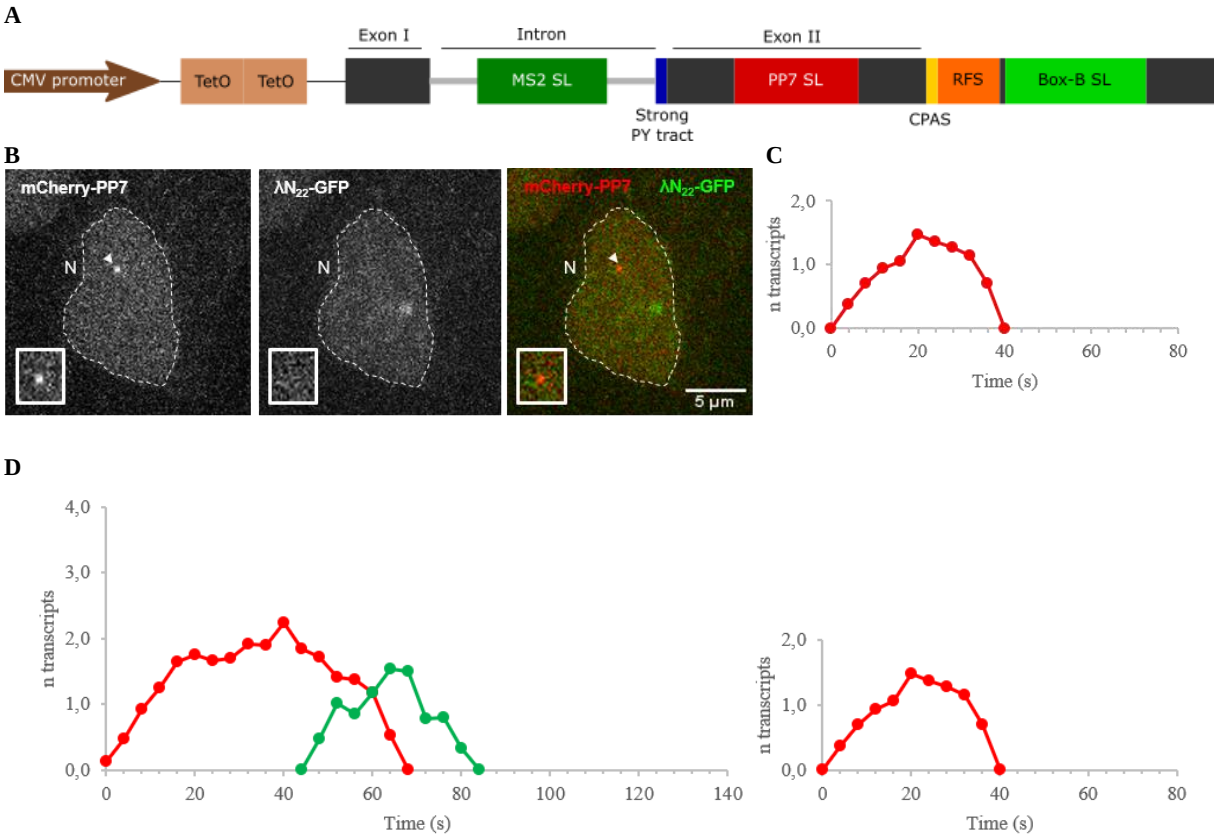


Figure 3.3 Effect of R-loop formation downstream the CPAS on transcription termination. **A** Representation of the IgM reporter gene IgM-1.7k-PYpA-baRFS. A sequence from the β-actin post-CPAS gene sequence which has been shown to form R-loops was inserted between the CPAS and the BoxB stem-loops. **B** Representative image of a HEK 293 cell, expressing the IgM-1.7k-PY-pA-baRFS. Transcripts labeled with mCherry-PP7 in exon II but not with λN_{22}-GFP were detected. The nucleus (N) is delimited with a dashed line and the TS of the reporter gene is indicated by an arrowhead. Insets depicting the diffraction limited objects in the highest intensity plane of acquired z-stacks. **C** A fluorescence intensity cycle representing a single transcript synthetized from the reporter gene with an RFS downstream the CPAS was plotted over time (s) in a line graph. **D** Comparison between a line plot showing the synthesis of more than one transcript from the reporter gene with a RFS downstream the CPAS, after induction of the expression of the reporter gene with a Dox concentration higher than 0.08 µg/ml and a line plot for the synthesis of only one transcript. The number (n) of transcripts was plotted over time (s).

To further confirm the influence of R-loop formation post-CPAS on RNA Pol II termination, a similar reporter gene (IgM-1.7k-PY-pA-NRFS) was engineered, by inserting a NRFS after the CPAS [66] (Fig. 3.4A). Contrary to our expectations, for single transcripts we only observed the synthesis of transcripts pre-CPAS, labeled with mCherry-PP7 (n=22) [Fig. 3.4B and C]. The time for transcription of the PP7 stem-loops in the reporter gene with a NRFS after the CPAS was measured at 14 ± 6 s and the corresponding calculated RNA Pol II speed was 5.85 kbp/min, which is approximate to the calculated value for the IgM-1.7k-PY and IgM-1.7k-PYpA-baRFS reporter genes. Considering these results, it appears that the presence of R-loops after this CPAS is not determinant for the efficiency of transcription termination in this reporter gene. Since the time calculated until the cleavage of the pre-CPAS transcript,

after RNA Pol II started transcribing the PP7 stem-loops sequence, was 35 ± 9 s, similar to IgM-1.7k-PY and IgM-1.7k-PYpA-baRFS reporter genes, the R-loop formation after this CPAS does not influence the efficiency of the cleavage in this reporter gene.

We also observed the effect of higher concentrations of Dox on the transcription of the reporter gene with a NRFS after the CPAS. As seen in Fig. 3.4D, with a Dox concentration of 0.08 µg/ml, the RNA Pol II transcribes the PP7 stem-loops, labeled with mCherry-PP7, and the BoxB stem-loops, labeled with λN_{22}-GFP, indicating readthrough of the polymerase. Once again, this result indicates that a high polymerase density influences the site and timing of transcription termination, leading to a delay in RNA polymerase release from the template and transcription of the sequence for BoxB stem-loops. Comparing the results obtained for the three reporter genes, when their expression is induced with a Dox concentration higher than 0.08 µg/ml, we detect the synthesis of a transcript post-CPAS in all cases. This result shows that independently of the R-loop formation after the CPAS, there is a delay in RNA Pol II release from the template and transcription of the sequence for BoxB stem-loops, due to a high polymerase density. The R-loop formation after the CPAS does not prevent polymerase readthrough when gene expression is high and cannot rescue immediate termination at the CPAS.

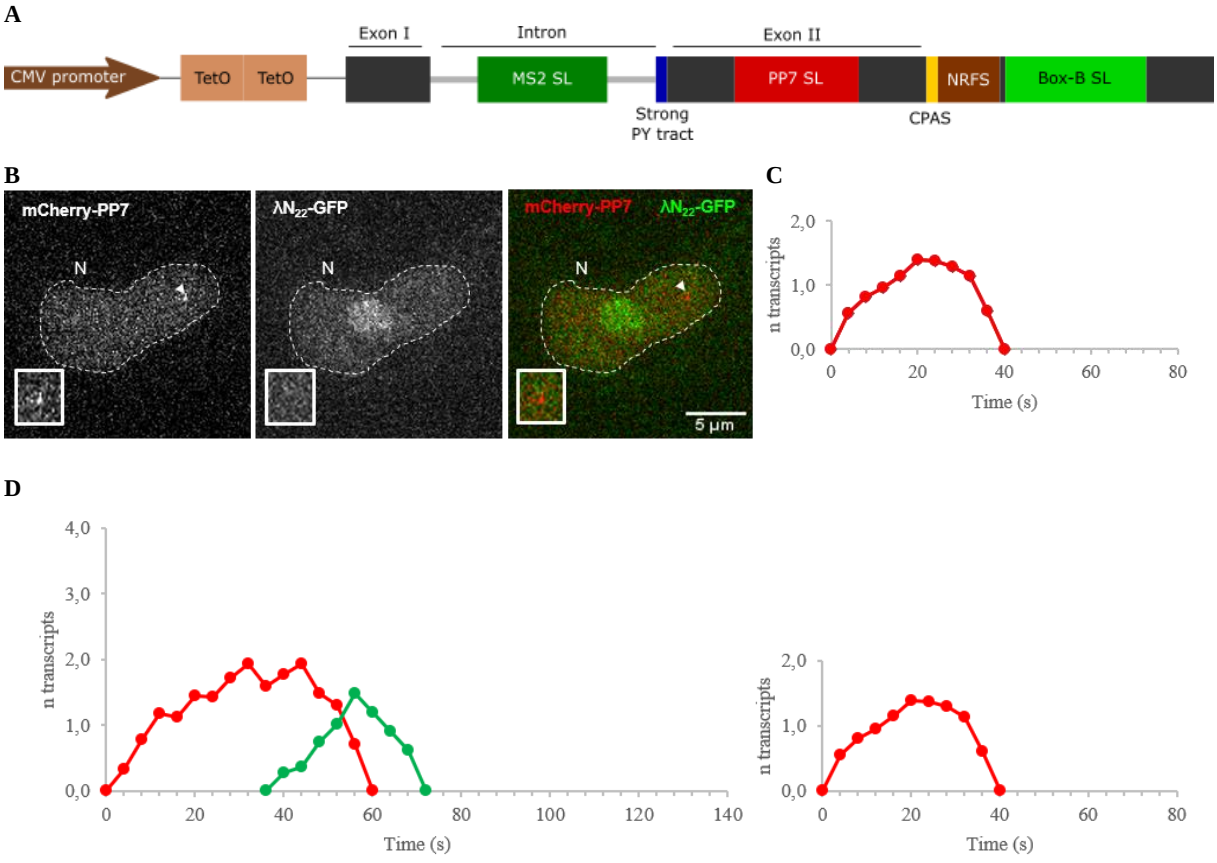


Figure 3.4 Transcription termination in the absence of R-loops downstream the CPAS. **A** Representation of the IgM-1.7k-PY-pA-NRFS reporter gene. A sequence that has been shown to not promote R-loop formation was inserted between the CPAS and the BoxB stem-loops. **B** Representative image of a HEK 293 cell, expressing the reporter gene IgM-1.7k-PY-pA-NRFS. Detection of transcripts tagged by the PP7 system in the exon II but no λN_{22}-GFP labeling post-CPAS. The nucleus (N) is delimited with a dashed line and the TS of the reporter gene is indicated in each cell with an arrowhead. Corresponding insets depicting the diffraction limited objects in the highest intensity plane of acquired z-stacks are shown below. **C** The number (n) of transcripts synthetized from the reporter gene with a NRFS downstream the CPAS was plotted over time (s) in line graphs, representing single transcript fluorescence intensity cycle over time graph. **D** Line plot of number (n) of transcripts against time (s). The one on the left shows the synthesis of more than one transcript from the reporter gene with a NRFS downstream the CPAS, after induction of the expression of the reporter gene with a Dox concentration higher than 0.08 µg/ml. The line plot on the right shows the synthesis of a single transcript, after induction of the expression of the reporter gene with a Dox concentration of 0.01 – 0.08 µg/ml.

3.4 Single-molecule sensitivity imaging of impaired transcription termination supports torpedo model

In the research field of transcription dynamics, there is still an ongoing debate about the exact mechanism of termination. It has mainly two opposing models, the allosteric model and the torpedo model. The allosteric model defends that, after transcribing the poly(A) signal, RNA Pol II undergoes a conformational change into a non-processive form and releases from the DNA template. On the other hand, the torpedo model postulates that an exoribonuclease degrades the transcript synthesized after the CPAS and displace the RNA Pol II from the DNA template when encounters it[39]. After determining the dynamics of RNA Pol II speeds and readthrough, we were interested to know if the measured transcription termination kinetics of the reporter genes could be explained by either model. To test if the torpedo model was able to explain our data, we reasoned that we should detect more readthrough after knocking down XRN2, the exoribonuclease responsible for degradation of post-CPAS transcripts and RNA Pol II termination.

To prove that XRN2 is involved in the termination of transcription of our reporter gene, this protein was knocked down by RNAi using a shRNA targeting XRN2. The knock-down efficiency was confirmed by western blot (Fig. 3.5A) and its effect on transcription termination was assessed by live cell imaging, using the previous system and settings and inducing the expression of the reporter gene with a Dox concentration of 0.01 – 0.08 µg/ml. After the knock-down of XRN2, the synthesis of single transcripts with labeling of exon II by mCherry-PP7 was observed in all the cells imaged (n=6), and the transcription of the region post-CPAS labeled with λN_{22}-GFP was detected in 33% of the cells imaged (Fig. 3.5B). Despite the low number of cells observed, this result shows that, in the absence of XRN2, there is an impairment in immediate CPAS transcription termination, leading RNA Pol II to transcribe after the CPAS. This set of data supports the torpedo model, since, according to this model, XRN2 degrades the post-CPAS transcript at the TS, when it is still connected to the template DNA via the RNA Pol II and, when XRN2 collides with the RNA Pol II, it is released from the DNA. We observed RNA Pol II transcribing the BoxB stem-loops after XRN2 knock-down, meaning that RNA Pol II is not released but readily transcribes if not forced by XRN2 to terminate just at the CPAS. This indicates that normally XRN2 degrades the uncapped nascent post-CPAS transcript very fast, before it can be detected with our system, and leads to a fast termination, when RNA Pol II density is low. In untreated cells the post-CPAS transcripts are not detected to be released in our system, only the mRNA. In an XRN2 knock-down, the mRNA release and the post-CPAS-RNA degradation and termination can be uncoupled. A similar uncoupling can be seen when RNA Pol II density is high at the reporter gene. However, to confirm these results, experimental adjustments and additional measurements are required in the future, because we have a low n=6 and the knock-down leads to an higher mortality among the cells, since XRN2 is essential for cells.

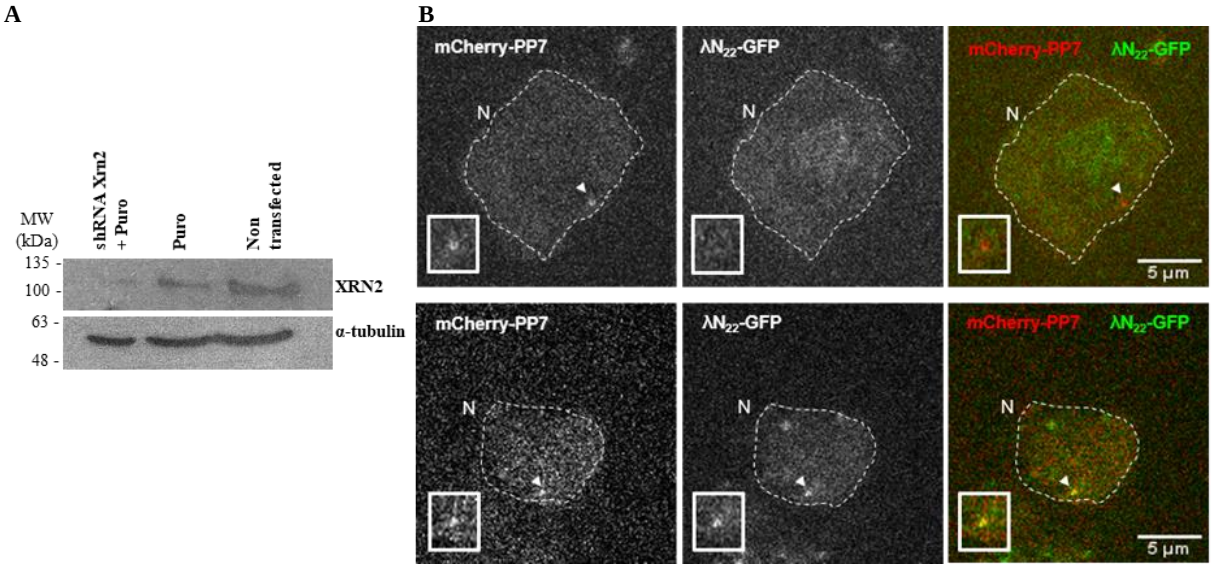


Figure 3.5 XRN2 knock-down supports the torpedo model for transcription termination. **A** Western blot confirming the knock-down of XRN2, using an RNAi technique. The positions of molecular weight (MW) markers are indicated on the left in kDa. The molecular weight of the XRN2 is 108.6 kDa and α-tubulin (49.9 kDa) was used as a loading. From left to right, in the first condition HEK 293 were transfected with shRNA targeting XRN2 and selected with 0.5 µg/ml Puromycin (Puro), next HEK 293 were only selected with 0.5 µg/ml Puro and, in the last condition, cells were neither transfected with shRNA targeting XRN2 or selected with Puro. **B** Representative image of 2 HEK 293 cells transcribing the IgM-1.7k-PY reporter gene after XRN2 knock-down. In the upper microscopy images, the TS is only labeled with m-Cherry-PP7 protein, as in 67% of the cases. In the bottom microscopy images, the TS is labeled with both mCherry-PP7 and λN_{22}-GFP proteins, representing 33% of the cells measured. The nucleus (N) is delimited with a dashed line and the TS with an arrowhead. Corresponding insets depicting the diffraction limited objects in the highest intensity plane of acquired z-stacks are shown below.

3.5 Splicing affects transcription termination

Previous research has already established that the pre-mRNA processing, specifically splicing, influences the cleavage/polyadenylation and termination of transcription, e.g. failure in splicing the last intron leads to failure in 3' end processing[95,96]. Therefore, we wanted to determine additional kinetic differences in the cleavage /termination process in our reporter genes when introducing mutations that impair splicing. For splicing to occur, a donor site at the 5' end of the intron, a branch site near the 3' end of the intron and an acceptor site at the 3' end of the intron are required. Most commonly, the splice donor site includes the dinucleotide GU at the 5' of the intron, within a less conserved region. The splice acceptor initiates the exon with the dinucleotide AG[97,98]. Of additional importance are the first nucleotides of the following exon with a G/AT consensus sequence. Upstream, there is a sequence rich in pyrimidines, called polypyrimidine (PY) tract, which is recognized by the spliceosome component U2AF and the polypyrimidine tract binding protein, promoting spliceosome assembly[99]. The strength of the PY tract, i.e., the efficient recognition of this site, depends on its constitution of pyrimidines, mainly uracils (in the RNA molecule), and the higher the pyrimidine percentage, the stronger the splice site.

To evaluate the impact of splice site strength on transcription termination efficiency, a reporter gene was constructed identical to the IgM-1.7k-PY reporter gene, although with a weak splice site in the 3' of the intron (IgM-1.7k-PYwsj). In Fig. 3.6A, are illustrated the strong PY tract from the IgM-1.7k-PY and a weaker PY tract from intron between exons C4 and M1 + 7 nt of M1 exon of IgM heavy chain constant region gene. As it has been shown that retention of introns leads to an impairment in cleavage/polyadenylation process and degradation of the unprocessed transcripts, we asked if we could detect changes in transcription readthrough and a possible slowing down or pausing of RNA Pol II when introns are not spliced. Because bulk methods may not detect these dynamic changes, our live cell imaging with single transcript detection might be able to do so. To measure the kinetics of cleavage/termination on unspliced reporter gene transcripts, we performed live cell microscopy

experiments. The HEK 293 cells with single copies of the reporter gene were transfected with the mCherry-PP7 and λN_{22}-GFP-encoding plasmids. Similarly, as before, cells were imaged in 3D over time. In this set of experiments, we detected production of transcripts pre-CPAS, labeled with, mCherry-PP7, and nascent transcripts after the CPAS, labeled with λN_{22}-GFP (Figure 3.6B and C). As seen before, the IgM-1.7k-PY reporter gene is efficiently processed. In contrast, the IgM-1.7k-PYwsj reporter shows a delay in cleavage of the pre-CPAS transcript. The value obtained was 41 ± 14 s, after RNA Pol II started transcribing the PP7 stem-loops, superior to the values obtained for the previous reporter genes, indicating that the retention of the intron also leads to an inefficient cleavage of the pre-CPAS transcript. Consequently, RNA Pol II is released from the template latter and readthrough is detected. The RNA Pol II takes 23 ± 10 s to transcribe the PP7 stem-loops, with a rate of 3.71 kbp/min, inferior to the value obtained for the transcription of the PP7 stem-loops in the IgM-1.7k-PY reporter gene. Although the polymerase takes the same time to transcribe BoxB stem-loops, they have a shorter length (0.999 kbp), so the RNA Pol II speed in that region is 2.34 kbp/min, much slower than for the transcription of PP7 stem-loops. This result indicates that RNA Pol II slows down when the splicing is not efficient, probably because of the unspliced intron, and slows down further in the CPAS region. We conclude that when splicing is not efficient, transcription termination is not immediate, with polymerase readthrough and slowdown.

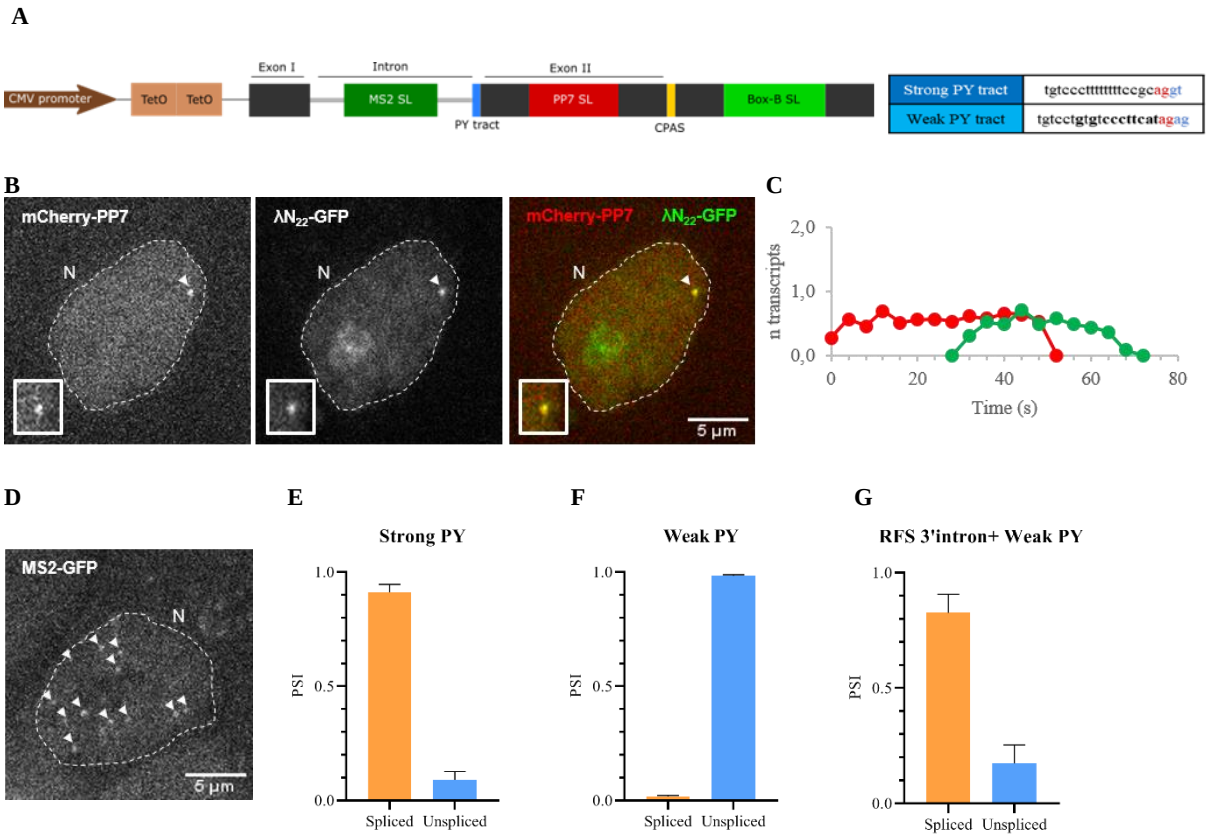


Figure 3.6 Influence of intron retention on mRNA cleavage and transcription termination. **A** Scheme of the reporter gene IgM-1.7k-PYwsj and comparison between the sequences of the strong and the weak PY tracts. The acceptor site at the 3' end of the intron is highlighted in red and the mutations in the weak PY tract are in bold. **B** Representative image of a HEK 293 cell, where the TS is labeled with both mCherry-PP7 and λN_{22}-GFP proteins. The nucleus (N) is delimited with a dashed line and the TS of the reporter gene is indicated by an arrowhead. Corresponding insets depicting the diffraction limited objects in the highest intensity plane of acquired z-stacks are shown below. **C** Representative single transcript fluorescence intensity cycle over time (s) graph, being n number (n) of transcripts synthetized from the reporter gene IgM-1.7k-PYwsj. **D** Representative image of a HEK 293 cell expressing the reporter gene IgM-1.7k-PYwsj. The introns are labeled with MS2-GFP. The arrowheads indicate unspliced transcripts dispersed in the nucleus (N). **E**, **F** and **G** Results of the q-PCR measurements to determine the relative amount of spliced and unspliced transcripts. For each condition the ratio of spliced transcripts was determined, and the PSI was calculated.

We imaged HEK 293 cells with single copies of the IgM-1.7k-PYwsj reporter gene to confirm if we could observe unspliced transcripts dispersed in the nucleus, labelling the intron with MS2-GFP. The mRNAs including the intron were only found in the nucleus but not in the cytoplasm, where they are probably degraded by the nuclear surveillance and degradation machinery[100] (Fig. 3.6D). To verify if the intron is included when the splice site is weak, we performed a qPCR, using the pairs of primers described in Annex 1. As expected, 91% of the transcripts synthetized from the IgM-1.7k-PY reporter gene were spliced, while 98% of the transcripts synthetized from the IgM-1.7k-PYwsj reporter gene were unspliced (Fig. 3.6E and F). The weak splice site is almost never recognized, leading to the inclusion of the intron in most of the mRNAs synthetized from this reporter gene.

3.6 Intronic R-loops upstream of 3' splice site rescue splicing and restore transcription termination efficiency

As demonstrated before, both splicing, cleavage and transcription termination are inefficient when there is a weak splice signal at the 3' end of the intron. However, some studies have already shown that the speed of polymerase transcription before the splice signal in the 3' end of the intron has an effect on the signal recognition[101,102]. They demonstrated that a slower transcription favors the inclusion of introns, despite the weak splice site. Other studies have indicated that RNA Pol II slows down at the CPAS site, due to R-loop formation, even though we were not able to measured that for our reporter gene by live cell measurement [45,51]. We hypothesized that the R-loop formation before a weak splice site could slow down RNA Pol II, providing more time for the spliceosome to recognize even weak 3' splice sites, therefore splicing of this intron could be more efficient. To validate this hypothesis, first we constructed the IgM-1.7k-3'intron-baRFS-PY reporter gene, introducing the same RFS from the β-actin gene 91bp upstream the weak PY tract (Fig. 3.7A). After generating a stable cell line by Flp-In recombination, we extracted the reporter gene (pre-)mRNA to perform a qPCR (n=3), in order to detect the ratio of spliced and unspliced transcripts, as described in Annex 3. Analyzing the production of transcripts from the reporter gene IgM-1.7k-3'intron-baRFS-PY, we surprisingly found that up to 83% of the transcripts were spliced (Fig. 3.6G), which represents a remarkable a rescue of the splicing, compared with the values obtained from the reporter gene IgM-1.7k-PYwsj, with 98% of transcripts unspliced. Thus, we can conclude that introducing the RFS in the 3' of the intron changes the splicing pattern of this particular reporter gene. We hypothesize that R-loop formation leads to the RNA Pol II slowdown just at the 3' splice site providing more time for the spliceosome to interact with the transcription complex and the 3' splice site to facilitate splicing.

As demonstrated before, if splicing is not efficient, then RNA Pol II termination is not efficient either. Since it was possible to rescue the splicing by inserting an RFS in the 3' end of the intron, we hypothesized that, transcription termination would be rescue too. HEK 293 cells with single copies of the reporter gene IgM-1.7k-3'intron-baRFS-PY were imaged (n=22) and, as hypothesized, it was detected in 82% of the cases only the signal for the exon II labeled with PP7-mCherry at the TS (Fig. 3.7B and C). Here, RNA Pol II is again released from the DNA template before transcribing the BoxB stem-loops. This result confirms earlier findings, that the transcription 3' end processing and termination are influenced by splicing or non-splicing of the last intron. In this case, the insertion of an RFS before the splice site has rescued splicing efficiency and, consequently, restored also partially the termination efficiency. We also detected in 18% of cases the synthesis of pre-CPAS and post-CPAS transcripts, the former labeled with mCherry-PP7 and the latter with λN_{22}-GFP (Fig. 3.7D and E). The kinetic parameters were calculated, as described before. We measured the time between the beginning of the transcription of the PP7 stem-loops and the cleavage, obtaining 32 ± 5 s when was detected only the synthesis of the pre-CPAS transcript and 37 ± 4 s, when readthrough was detected. This result shows that the insertion of an RFS before the splice site has also restored partially the cleavage efficiency. The

RNA Pol II takes 15 ± 5 s to transcribe PP7 stem-loops, at 5.58 kbp/min, when only the transcripts labeled with mCherry-PP7 are synthetized. However, when BoxB stem-loops are also transcribed, RNA Pol II takes 23 ± 2 s, corresponding to a transcription rate of 3.71 kbp/min. In this situation, the BoxB stem-loops are transcribed in 5 ± 2 s, at a transcription rate of 11.24 kbp/min. One explanation for this could be that the 0.999 kb BoxB sequence is not completely transcribed and the nascent transcript is degraded before RNA Pol II transcribes all the stem-loop sequence. Therefore, the value measured derived from the time of increase of fluorescence signal would be overestimated. On the other hand, the readthrough in the transcription of the reporter gene with R-loops in 3' end of the intron was detected in only 18% of the imaged cells (n = 3), which is a very low number of examples to precisely calculate the speed of RNA Pol II after CPAS.

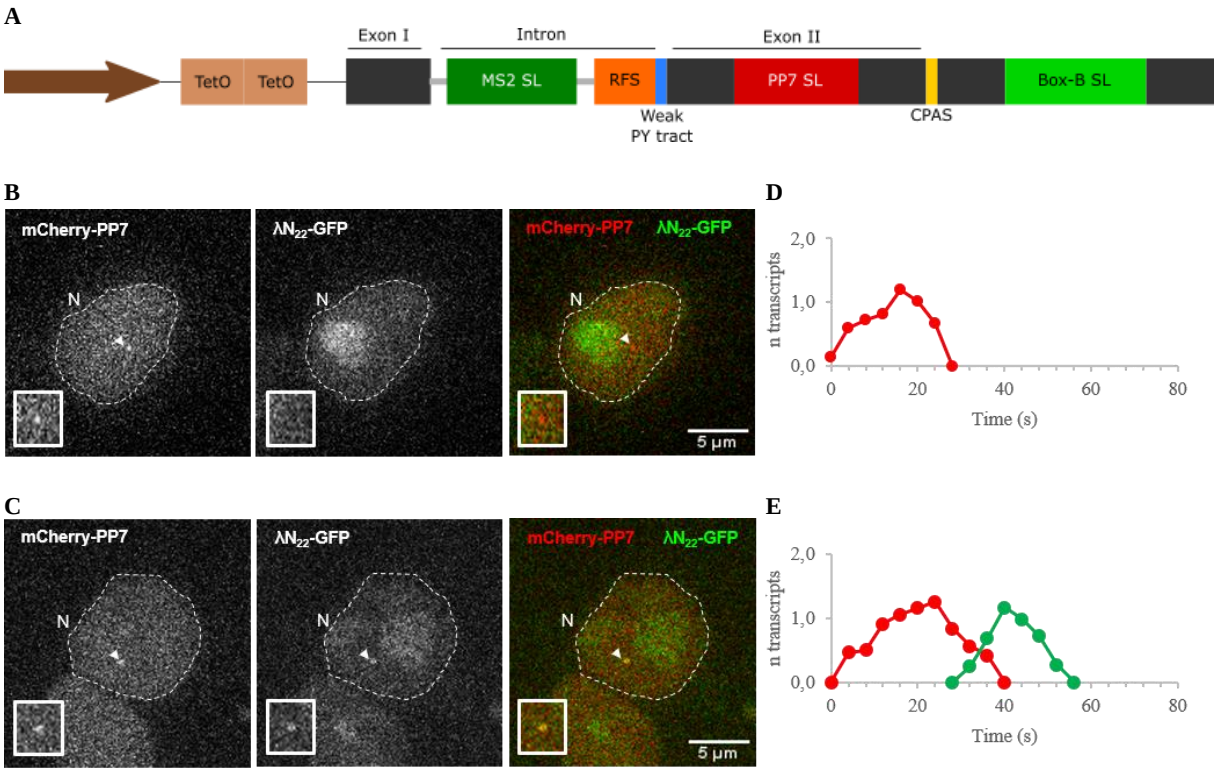


Figure 3.7 The role of R-loops in the rescue of splicing, cleavage and termination efficiency. **A** Scheme of the reporter gene IgM-1.7k-3'intron-baRFS-PY indicating the position of the RFS from β-actin gene introduced just upstream of the 3' splice site. **B** and **C** Representative image of 2 HEK 293 cells transcribing the reporter gene with an RFS in the 3' end of the intron. The nucleus (N) is delimited with a dashed line and the TS of the reporter gene is indicated in each cell with an arrowhead. Corresponding insets depicting the diffraction limited objects in the highest intensity plane of acquired z-stacks are shown below. **B** In 82% of the cells imaged, there were detected only transcripts tagged by the PP7 system in the exon II. In **C** TS is labeled with both mCherry-PP7 and λN_{22}-GFP proteins. This cell is representative of the 18% of the cases where both labels were detected in the TS. **D** and **E** Line plots depicting a complete cycle of fluorescence gain and loss present in a time-lapse series. The calibrated number of transcripts (n) was plotted over time (s). **D** A cycle of a transcription representing 82% of the cases with the TS labeled only with mCherry-PP7. **E** A cycle of a transcript representing the 18% of cells with the TS-labeled with mCherry-PP7 and λN_{22}-GFP.

To test if the 25xBoxB stem-loop sequence post-CPAS is transcribed completely or only partially, we used another reporter gene for calibration of the signal intensity. In the IgM-1.7k-PY-exonII-BoxB reporter gene version, the 25xBoxB sequence was inserted in the exon II 3' UTR at the same position where the 24xPP7 stem-loop sequence was introduced in the other reporter genes (Fig. 3.8). Here, the 25xBoxB sequence will be completely transcribed after induction of the expression of this new reporter gene and the signal intensity measured for single transcripts should represent labeling of the full length 25xBoxB stem-loop sequence (Annex 1). Comparing the calculated average for the maximum TFI value corresponding to the transcription of 25xBoxB sequence, labeled with λN_{22}-GFP, the value obtained for

the transcription of IgM-1.7k-3'intron-baRFS-PY reporter gene is between the values obtained for the IgM-1.7k-PY-exonII-BoxB and IgM-1.7k-PY reporter genes, under overexpression of RNaseH1, suggesting that BoxB stem-loops are completely transcribed for the former case (Table 1). However, despite the unrealistic high value for the RNA Pol II speed over the BoxB stem-loops in the IgM-1.7k-3'intron-baRFS-PY reporter gene, we are not able to conclude if, in these case, the BoxB stem-loops are completely transcribed, due to the low number of examples acquired.

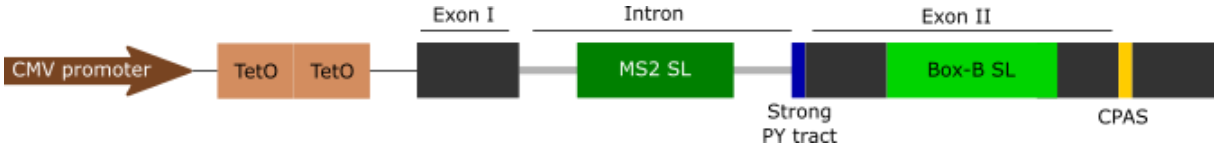


Figure 3.8 Scheme of the IgM reporter gene with binding sites for λN_{22}-GFP in exon II. The interaction of fluorescently tagged RNA stem-loop binding proteins with the respective stem-loops allows the visualization of the RNAs.

Table 3.1 Comparison of the calculated average for the maximum TFI value corresponding to the transcription of 25xBoxB sequence between the reporter gene IgM-1.7k-PY-exonII-BoxB, the reporter gene IgM-1.7k-PY, under overexpression of RNaseH1, and the reporter gene IgM-1.7k-PY-3'intron-baRFS.

	25xBoxB in exon II	25xBoxB post-CPAS IgM-1.7k-PY reporter, under overexpression of RNaseH1	25xBoxB post-CPAS IgM-1.7k-intron-3´RFS-PY
mean max. TFI value (± StDev)	55 ± 21	64 ± 11	60 ± 8

In sum, we demonstrated in this project that when RNAseH is overexpressed, the transcription termination is not immediate in 27% of the cell imaged, being possible to detect the readthrough of the polymerase and an higher transcription rate upstream the CPAS. Further, we show that R-loop formation after the CPAS is not essential for an immediate transcription termination, by comparing the IgM-1.7k-PYpA-baRFS and IgM-1.7k-PYpA-NRFS reporter genes. We also demonstrate an evidence for the torpedo model by observing a non-immediate termination, in 33% of the cases, when XRN 2 is knocked down. Moreover, we show that the retention of the intron in the mRNA leads to impairment of the pre-CPAS transcript cleavage. However, the partial rescue of splicing efficiency by inserting an RFS upstream the PY1 tract, leads to efficient cleavage of the mRNA and an immediate transcription termination in 82% of the cells imaged.

4. Conclusions and Discussion

Transcription is the first step of gene expression, consisting in the synthesis of an RNA molecule complementary to a DNA template, catalyzed by an RNA polymerase. The type of transcription addressed in this project is the one catalyzed by RNA Pol II to synthetize mRNA. The transcription by RNA Pol II is tightly regulated, for example, by binding of transcription factors to the polymerase, epigenetic modifications or formation of secondary structures[1,6]. R-loops are by-products of transcription that also have been found to have a role in transcription regulation. They result from the hybridization of the RNA molecule produced by RNA Pol II with the complementary DNA strand, however how they affect transcription termination and mRNA processing is not well understood[32,40,41]. In this project we investigated the kinetics of the pre-mRNA 3' end processing and the influence of R-loop formation on the process of transcription and RNA processing, namely the cleavage/polyadenylation, transcription termination and splicing. To achieve the proposed objectives, we used spinning disk confocal microscopy with single molecule sensitivity, which allows to follow the production and release/degradation of single transcripts in real-time, in live cells and study the kinetic parameters of RNA Pol II during transcription.

One of the objectives of this project was to provide kinetic description of the pre-mRNA 3' end processing and transcription termination. In the first experiments we observed that, during the transcription of the IgM-1.7k-PY reporter gene, the PP7 stem-loops sequence in exon II is transcribed at a rate of around 4.94 kbp/min and the transcription of the BoxB stem-loops post-CPAS is not detected. We conclude that RNA Pol II releases from the DNA before transcribing the 25xBoxB sequence or the nascent transcript is immediately degraded, most likely, by XRN2[39], i.e., transcription termination is immediate after reaching the CPAS. The live-imaging experiments allow us to follow the transcription of the sequences codifying for stem-loops in real-time, only with an interval of 4 s, and the specificity of the viral proteins fused with fluorescent to the repetitive stem-loops enables the detection of single-transcripts, amplifying the signal. However, when the Dox concentration used to induce the transcription of the reporter gene was higher than 0.08 µg/ml, there was the synthesis of multiple transcripts simultaneously and readthrough was detected. Concentrations of Dox higher than 0.08 µg/ml increase the polymerase density in the reporter gene and transcription after the CPAS is detected through the binding of λN_{22}-GFP to the BoxB stem-loops. This indicates that a high polymerase density influences the site and timing of transcription termination, leading to a delay in RNA polymerase release from the template and transcription of the BoxB stem-loops sequence. In fact, a previous study demonstrated that a high RNA Pol II density at a gene leads to an increase in elongation rate[89]. As there was the simultaneous production of several transcripts, it was impossible to define the beginning and the end of transcription of each transcript. Therefore, the RNA Pol II speed was not calculated when multiple transcripts were synthetized.

Previous studies showed that R-loops could play a role in transcription termination by inducing pausing or slowdown of RNA Pol II and contributing to an immediate transcription termination[66–68]. For that reason and once the sequence right after the CPAS in our reporter gene is prone to the formation of R-loops, we tested the influence of RNaseH1, an enzyme which degrades R-loops, in the efficiency of transcription termination of our reporter gene. By live cell imaging, we verify that, in 27% of the cells transfected with RNAseH1-GR-iRFP, the BoxB stem-loop sequence was transcribed, meaning that RNA Pol II keeps transcribing beyond the CPAS and the transcription termination is not immediate. The analysis of the RNA Pol II speed shows that the enzyme is faster transcribing the PP7 stem-loops sequence when readthrough was detected, around 6.82 kbp/min. These results led us to conclude that probably RNaseH1 is degrading R-loops which form along the reporter gene and help to control the RNA Pol II speed. So, in 27% of the cases, as the R-loops are degraded, the RNA Pol II speed increases

and the enzyme is not able to leave the DNA before transcribing the BoxB stem-loops. Thus, we hypothesize that the R-loop formation along the gene is important to control RNA Pol II speed and that a velocity of the polymerase significatively higher than 6.0 kbp/min leads to an non-immediate transcription termination and consequent readthrough, as already shown in a previous study[25]. In this case, the detection of transcription after the CPAS can be due to the difficulty for the XRNA2 to catch RNA Pol II, once it is much faster. However, the measurements are not precise, so, the results obtained may be due to other side effects of RNaseH1, for example at the promoter or at the spliceosome.

To address the question if the R-loop formation after the CPAS influences transcription termination efficiency, we used two similar reporter genes, IgM-1.7k-PY-pA-baRFS and IgM-1.7k-PY-pA-NRFS. Only the transcription of the PP7 stem-loops sequence was observed in both cases, meaning that there was not readthrough and the transcription termination is immediate in the presence or in the absence of R-loops after the CPAS. So, we can conclude that the R-loop formation after the CPAS is not important for the transcription termination in this gene, despite the evidences previously presented for other genes[51,66,103]. In fact, there are also other genes where R-loop formation after the CPAS does not play a role in cleavage and transcription termination, as the *SNRPN* [104]. The velocity calculated for the transcription of the PP7 stem-loops sequence for each reporter gene (6.04 kbp/min for the IgM-1.7k-PY-pA-baRFS reporter gene and 5.85 kbp/min for the IgM-1.7k-PY-pA-NRFS reporter gene) are considered similar to the value calculated for the IgM-1.7k-PY reporter gene because the cells are imaged only in 4s interval and the average time measured for the transcription of this region in these reporter genes (14 s) is close to the timepoint 16 s of the timelapse, as the one measured for the IgM-1.7k-PY reporter gene (17 s). So, there is not a significant difference ($p>0.05$, t-test) between the speed of RNA pol II in the transcription of the PP7 stem-loops sequence measured for the three reporter genes. Despite the fact that we did not detected a impairment in transcription termination when there is no R-loop formation downstream the CPAS, this does not invalidate the torpedo model in this case because, XRN2 is recruited by the multifunctional protein p54nrb/PSF [105], not depending on R-loops.

As explained before, there are two models for RNA Pol II transcription termination: the allosteric model and the torpedo model. The former defends that RNA Pol II is transformed by a conformational change into a non-processive form after transcribing the CPAS, releasing the DNA template, while the latter posits that a exoribonuclease degrades the transcript synthesized after the CPAS and displace the RNA Pol II when encounters it[39]. Then, we inquire which transcription termination model could explain the termination process in the IgM-1.7k-PY reporter gene. To accomplish that, we knocked-down XRN2, which is a fundamental protein for the degradation of the nascent transcript after the transcription of the CPAS, according to the torpedo model. We detected readthrough in 33% of the cells imaged meaning that, in these cases, the XRN2 is essential for an efficient transcription termination. These results are an evidence of the torpedo model for the transcription termination in this reporter gene. Further, in untreated cells we only detected mRNA to be released in our system, and not post-CPAS transcripts. Thus, the mRNA release and the post-CPAS-RNA degradation and termination can be uncoupled by the XRN2 knockdown, or when Pol II density is high at the reporter gene, as seen before. Nevertheless, it is necessary to repeat the experiment, since the cells transfected with the plasmid for the shRNA seemed stressed and exhibited morphologic alterations, meaning that XRN2 is important for the cell survival[106]. Moreover, despite the confirmation of the XRN2 knock down by western blot, in live cell microscopy experiments, we cannot ensure that XRN2 is knocked down in that specific cell. To have a more accurate result for the percentage of the cases with readthrough when XRN2 is knocked down, we could perform live cell imaging in a 35 mm petri dishes with 10 mm glass bottom with a grid, fix the cells and perform immunofluorescence to confirm if there was the knock down of XRN2 in cells imaged before.

Splicing is also an important cellular process that consists in the processing of the pre-mRNA, by removing the non-coding intron sequences, and is mostly co-transcriptional. In fact, transcription and

splicing are coupled. Some studies have already shown that splicing can influence cleavage and polyadenylation of the pre-mRNA and transcription termination[95,96]. So, we evaluated if the impairment of splicing affects the kinetics of the cleavage of the transcript and transcription termination. Imaging cells expressing the reporter gene IgM-1.7k-PYwsj, we detected the transcription of a sequence downstream of the CPAS, indicating a readthrough and a non-immediate transcription termination. Considering the kinetic parameters calculated, we observed a delay in the mRNA cleavage (41 ± 14 s), comparing to the IgM-1.7k-PY reporter gene (35 ± 10 s). The RNA Pol II was also slower over the PP7 stem-loops (3.71 kbp/min). Perhaps the spliceosome still bound to the RNA Pol II avoiding the binding of the factors needed for cleavage and transcription termination. Since the mRNA was not cleaved, there is not substrate for XRN2 (5' end of the nascent transcript after the transcription of the CPAS), and the exoribonuclease cannot reach the RNA Pol II, to release it from the DNA template, confirming the torpedo model. Thus, we conclude that if splicing is not efficient, the cleavage is not efficient, and the transcription termination is not immediate. Indeed, Brody *et al.* (2011) have already demonstrated that the RNA Pol II is retained in association with the chromatin when splicing is not efficient[107].

Some studies showed that R-loops play an important role in regulating transcription[43]. So, we hypothesized that R-loops could also influenced splicing efficiency. To evaluate the influence of R-loops in splicing a qPCR was performed to detect the type of transcripts synthetized from three reporter genes: IgM-1.7k-PY, IgM-1.7k-PYwsj and IgM-1.7k-3'intron-baRFS-PY. As shown before[81], 91% of the transcripts synthetized from the IgM-1.7k-PY reporter gene were spliced, while 98% of the transcripts produced from the IgM-1.7k-PYwsj reporter gene were unspliced. In addition, we demonstrated that the R-loop formation before a weak PY tract in the IgM-1.7k-3'intron-baRFS-PY reporter gene leads to the rescue of splicing efficiency, since the detection of spliced transcripts increased (83%), comparing with the similar reporter gene without the RFS before the weak PY tract. Probably, the presence of R-loops before the weak PY tract induces the slowdown of RNA Pol II, providing more time to the spliceosome to recognize this weak splice site, enhancing the splicing efficiency.

Once the efficiency of the splicing was restored because of the R-loop formation at the 3' end of the intron, and the efficiency of splicing showed before to influence transcription termination, we also tested how R-loop formation in the 3' end of the intron affected transcription termination. By live cell microscopy, we detected only the transcription of the PP7 stem-loops in 82% of the cells imaged. We demonstrated that R-loop formation before a weak PY track can restore partially not only splicing efficiency but also the efficiency of transcription termination. Indeed, comparing the transcription of the IgM-1.7k-PYwsj reporter gene with the IgM-1.7k-3'intron-baRFS-PY reporter gene, there was a significant decrease in the readthrough, after the cleavage of the pre-mRNA. Hence, splicing and transcription termination are connected and impairment of splicing leads also to an impairment in transcription termination.

Although this project contributed to elucidate the R-loops function in the process of transcription and RNA processing, some questions remain to answer. Regarding the influence of R-loops in splicing, we could also evaluate the kinetics of splicing, by measuring the lifetime of the intron in the transcript, using live cell imaging. The reporter genes used in this study have got an array of MS2 stem-loops placed into the intron. So, MS2 protein can be fused with a fluorophore and used to recognize these stem-loops in the transcript. Measuring the cycles of fluorescence gain and loss that started and returned to background level, it is possible to infer the lifetime of the intron and evaluate if it is affected by the presence of R-loops in the 3' end of the intron. Some studies also indicate that R-loops could have a role in transcription initiation, therefore, we could evaluate the effect of R-loop formation on the initiation of transcription, using the same method.

To summarize, in this project, we demonstrated that when RNaseH1 is overexpressed, the rate of transcription of RNA Pol II is abnormally high, leading to readthrough. We also showed that R-loops formation after the CPAS is not determinant for the transcription termination. By the knock down of XRN2 we showed that this enzyme contributes to an immediate transcription termination. We also demonstrated the presence of an intron in the nascent transcript influences the cleavage of mRNA and transcription termination efficiency. The R-loop formation in the 3' end of the intron revealed to rescue the splicing efficiency when the signal for splicing is weak and, consequently, the cleavage and transcription termination are also efficient, due to the fact that there is no intron in the transcript.

5. References

1. Thomas, M. C. & Chiang, C. M. The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41**, 105–178 (2006).
2. Cramer, P. *et al.* Structure of Eukaryotic RNA Polymerases. *Annu. Rev. Biophys.* **37**, 337–352 (2008).
3. Dieci, G., Conti, A., Pagano, A. & Carnevali, D. Identification of RNA polymerase III-transcribed genes in eukaryotic genomes. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1829**, 296–305 (2013).
4. Moss, T. & Stefanovsky, V. Y. At the Center of Eukaryotic Life. *Cell* **109**, 545–548 (2002).
5. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
6. Shandilya, J. & Roberts, S. G. E. The transcription cycle in eukaryotes: From productive initiation to RNA polymerase II recycling. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1819**, 391–400 (2012).
7. Zhou, Q., Li, T. & Price, D. H. RNA Polymerase II Elongation Control. *Annu. Rev. Biochem.* **81**, 119–143 (2012).
8. Matsui, T., Segall, J., Weil, P. & Roeder, R. Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. *J Biol Chem* **255**, 11992–6 (1980).
9. Starr, D. B. & Hawley, D. K. TFIID binds in the minor groove of the TATA box. *Cell* **67**, 1231–1240 (1991).
10. Cadenas, D. L. & Dahmus, M. E. Messenger RNA Synthesis in Mammalian Cells Is Catalyzed by the Phosphorylated Form of RNA Polymerase II. **262**, 12468–12474 (1987).
11. Heidemann, M., Hintermair, C., Voß, K. & Eick, D. Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1829**, 55–62 (2013).
12. Wang, W., Carey, M. & Gralla, J. Polymerase II promoter activation: closed complex formation and ATP-driven start site opening. *Science (80-.).* **255**, 450–453 (1992).
13. Ping, Y.-H. & Rana, T. M. DSIF and NELF Interact with RNA Polymerase II Elongation Complex and HIV-1 Tat Stimulates P-TEFb-mediated Phosphorylation of RNA Polymerase II and DSIF during Transcription Elongation. *J. Biol. Chem.* **276**, 12951–12958 (2001).
14. Larochelle, S. *et al.* Cyclin-dependent kinase control of the initiation-to-elongation switch of RNA polymerase II. *Nat. Struct. Mol. Biol.* **19**, 1108–1115 (2012).
15. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).
16. Kelleher, R. J., Flanagan, P. M. & Kornberg, R. D. A novel mediator between activator proteins and the RNA polymerase II transcription apparatus. *Cell* **61**, 1209–1215 (1990).
17. Lorch, Y., Zhang, M. & Kornberg, R. D. RSC Unravels the Nucleosome. *Mol. Cell* **7**, 89–95 (2001).
18. Georgakopoulos, T. & Thireos, G. Two distinct yeast transcriptional activators require the function of the GCN5 protein to promote normal levels of transcription. *EMBO J.* **11**, 4145–52 (1992).
19. Jülicher, F. & Bruinsma, R. Motion of RNA Polymerase along DNA: A Stochastic Model. *Biophys. J.* **74**, 1169–1185 (1998).
20. Chen, F. X., Smith, E. R. & Shilatifard, A. Born to run: Control of transcription elongation by

RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **19**, 464–478 (2018).

21. Sakharkar, M. K., Chow, V. T. K. & Kangueane, P. Distributions of Exons and Introns in the Human Genome. *In Silico Biol.* **4**, 387–393 (2004).

22. Proudfoot, N. J., Furger, A., Dye Sir, M. J. & Dunn, W. Review Integrating mRNA Processing with Transcription. *Cell* **108**, 501–512 (2002).

23. Keller, W. The RNA lariat: A new ring to the splicing of mRNA precursors. *Cell* **39**, 423–425 (1984).

24. Beyer, A. L. & Osheim, Y. N. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes & Dev.* **2**, 754–765 (1988).

25. Fong, N. *et al.* Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Mol. Cell* **60**, 256–267 (2015).

26. Sanford, J. R. Pre-mRNA splicing: life at the centre of the central dogma. *J. Cell Sci.* **117**, 6261–6263 (2004).

27. Hang, J., Wan, R., Yan, C. & Shi, Y. Structural basis of pre-mRNA splicing. *Science (80-.).* **349**, 1191–1198 (2015).

28. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136**, 701–718 (2009).

29. Yan, C. *et al.* Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science (80-.).* **349**, 1182–1191 (2015).

30. Kuehner, J. N., Pearson, E. L. & Moore, C. Unravelling the means to an end: RNA polymerase II transcription termination. *Nat. Rev. Mol. Cell Biol.* **12**, 283–294 (2011).

31. Richard, P. & Manley, J. L. Transcription termination by nuclear RNA polymerases. *Genes Dev.* **23**, 1247–69 (2009).

32. Nick J. Proudfoot. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science (80-.).* **352**, (2016).

33. Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.* **16**, 190–202 (2015).

34. Porrua, O., Boudvillain, M. & Libri, D. Transcription Termination: Variations on Common Themes. *Trends Genet.* **32**, 508–522 (2016).

35. Ustyantsev, I. G., Golubchikova, J. S., Borodulina, O. R. & Kramerov, D. A. Canonical and noncanonical RNA polyadenylation. *Mol. Biol. (Mosk).* **51**, 262–273 (2017).

36. Kamieniarz-Gdula, K. *et al.* Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Mol. Cell* **74**, 158-172.e9 (2019).

37. Zhang, Z. CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11. *Genes Dev.* **19**, 1572–1580 (2005).

38. Gilmour, D. S. & Fan, R. Derailing the Locomotive: Transcription Termination. *J. Biol. Chem.* **283**, 661–664 (2008).

39. Luo, W. & Bentley, D. A ribonucleolytic rat torpedoes RNA polymerase II. *Cell* **119**, 911–914 (2004).

40. Crossley, M. P., Bocek, M. & Cimprich, K. A. R-Loops as Cellular Regulators and Genomic Threats. *Mol. Cell* **73**, 398–411 (2019).

41. Aguilera, A. & García-Muse, T. R Loops: From Transcription Byproducts to Threats to Genome Stability. *Mol. Cell* **46**, 115–124 (2012).

42. Drolet, M. *et al.* Overexpression of RNase H partially complements the growth defect of an Escherichia coli delta topA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I. *Proc. Natl. Acad. Sci.* **92**, 3526–3530 (1995).

43. Santos-Pereira, J. M. & Aguilera, A. R loops: New modulators of genome dynamics and function. *Nat. Rev. Genet.* **16**, 583–597 (2015).

44. Westover, K. D. Structural Basis of Transcription: Separation of RNA from DNA by RNA Polymerase II. *Science (80-.).* **303**, 1014–1016 (2004).

45. Skourti-Stathaki, K. A double-edged sword : R loops as t h reat s t o gen om e in t egrit y an d pow erfu l regu lat ors of gen e expression. *Genes Dev.* 1384–1396 (2014). doi:10.1101/gad.242990.114.Freely

46. Shaw, N. N. & Arya, D. P. Recognition of the unique structure of DNA:RNA hybrids. *Biochimie* **90**, 1026–1039 (2008).

47. Chédin, F. Nascent Connections: R-Loops and Chromatin Patterning. *Trends Genet.* **32**, 828–838 (2016).

48. De Magis, A. *et al.* DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proc. Natl. Acad. Sci.* **116**, 816–825 (2019).

49. Chon, H. *et al.* Contributions of the two accessory subunits, RNASEH2B and RNASEH2C, to the activity and properties of the human RNase H2 complex. *Nucleic Acids Res.* **37**, 96–110 (2009).

50. Groh, M., Albulescu, L. O., Cristini, A. & Gromak, N. Senataxin: Genome Guardian at the Interface of Transcription and Neurodegeneration. *J. Mol. Biol.* **429**, 3181–3195 (2017).

51. Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. *Mol. Cell* **42**, 794–805 (2011).

52. Manzo, S. G. *et al.* DNA Topoisomerase I differentially modulates R-loops across the human genome. *Genome Biol.* **19**, 100 (2018).

53. Wahba, L. & Koshland, D. The Rs of Biology: R-Loops and the Regulation of Regulators. *Mol. Cell* **50**, 611–612 (2013).

54. Wahba, L., Gore, S. K. & Koshland, D. The homologous recombination machinery modulates the formation of RNA-DNA hybrids and associated chromosome instability. *Elife* **2013**, 1–20 (2013).

55. Tan-Wong, S. M. & Proudfoot, N. J. Rad51, friend or foe? *Elife* **2**, (2013).

56. Toriumi, K., Tsukahara, T. & Hanai, R. R-Loop Formation In Trans at an AGGAG Repeat. *J. Nucleic Acids* **2013**, 1–7 (2013).

57. Illingworth, R. S. & Bird, A. P. CpG islands - 'A rough guide'. *FEBS Lett.* **583**, 1713–1720 (2009).

58. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *PNAS* **103**, 1412–1417 (2005).

59. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).

60. Payer, B. & Lee, J. T. X Chromosome Dosage Compensation: How Mammals Keep the Balance. *Annu. Rev. Genet.* **42**, 733–772 (2008).

61. Guibert, S., Forné, T. & Weber, M. Dynamic regulation of DNA methylation during mammalian development. *Epigenomics* **1**, 81–98 (2009).

62. Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I. & Chédin, F. R-Loop Formation Is a

Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Mol. Cell* **45**, 814–825 (2012).

63. Xu, Z., Zan, H., Pone, E. J., Mai, T. & Casali, P. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nat. Rev. Immunol.* **12**, 517–531 (2012).

64. Yu, K., Chedin, F., Hsieh, C.-L., Wilson, T. E. & Lieber, M. R. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat. Immunol.* **4**, 442–451 (2003).

65. Muramatsu, M. *et al.* Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. *Cell* **102**, 553–563 (2000).

66. Skourti-Stathaki, K., Kamieniarz-Gdula, K. & Proudfoot, N. J. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* **516**, 436–439 (2014).

67. Huppert, J. L., Bugaut, A., Kumari, S. & Balasubramanian, S. G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.* **36**, 6260–6268 (2008).

68. Ginno, P. A., Lim, Y. W., Lott, P. L., Korf, I. & Chédin, F. GC skew at the 5’ and 3’ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res.* **23**, 1590–1600 (2013).

69. Gomez-Gonzalez, B., Felipe-Abrio, I. & Aguilera, A. The S-Phase Checkpoint Is Required To Respond to R-Loops Accumulated in THO Mutants. *Mol. Cell. Biol.* **29**, 5203–5213 (2009).

70. Gan, W. *et al.* R-loop-mediated genomic instability is caused by impairment of replication fork progression. *Genes Dev.* **25**, 2041–2056 (2011).

71. Costantino, L. & Koshland, D. The Yin and Yang of R-loop biology. *Curr. Opin. Cell Biol.* **34**, 39–45 (2015).

72. Sollier, J. & Cimprich, K. A. Breaking bad: R-loops and genome integrity. *Trends Cell Biol.* **25**, 514–522 (2015).

73. Sollier, J. *et al.* Transcription-Coupled Nucleotide Excision Repair Factors Promote R-Loop-Induced Genome Instability. *Mol. Cell* **56**, 777–785 (2014).

74. García-Muse, T. & Aguilera, A. Transcription–replication conflicts: how they occur and how they are resolved. *Nat. Rev. Mol. Cell Biol.* **17**, 553–563 (2016).

75. Helmrich, A., Ballarino, M., Nudler, E. & Tora, L. Transcription-replication encounters, consequences and genomic instability. *Nat. Struct. Mol. Biol.* **20**, 412–418 (2013).

76. Wang, J., Haeusler, A. R. & Simko, E. A. Emerging role of rna-dna hybrids in c9orf72-linked neurodegeneration. *Cell Cycle* **14**, 526–532 (2015).

77. Haeusler, A. R., Donnelly, C. J. & Rothstein, J. D. The expanding biology of the C9orf72 nucleotide repeat expansion in neurodegenerative disease. *Nat. Rev. Neurosci.* **17**, 383–395 (2016).

78. Coleman, R. A. *et al.* Imaging transcription: Past, present, and future. *Cold Spring Harb. Symp. Quant. Biol.* **80**, 1–8 (2016).

79. Sako, Y. & Yanagida, T. Single-molecule visualization in cell biology. *Nat. Rev. Mol. Cell Biol.* **Suppl**, SS1-5 (2003).

80. Zlatanova, J. & van Holde, K. Single-Molecule Biology: What Is It and How Does It Work? *Mol. Cell* **24**, 317–329 (2006).

81. Martin, R. M., Rino, J., Carvalho, C., Kirchhausen, T. & Carmo-Fonseca, M. Live-Cell Visualization of Pre-mRNA Splicing with Single-Molecule Sensitivity. *Cell Rep.* **4**, 1144–1155 (2013).

82. Gossen, M. & Bujard, H. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 5547–5551 (1992).

83. Daigle, N. & Ellenberg, J. λN-GFP: an RNA reporter system for live-cell imaging. *Nat. Methods* **4**, 633–636 (2007).

84. Martin, R. M., Rino, J., de Jesus, A. C. & Carmo-Fonseca, M. Single-Molecule Live-Cell Visualization of Pre-mRNA Splicing. in *Post Transcriptional Gene Regulation* 335–350 (2016). doi:10.1007/978-1-4939-3067-8_22

85. Rino, J., de Jesus, A. C. & Carmo-Fonseca, M. STaQTool: Spot tracking and quantification tool for monitoring splicing of single pre-mRNA molecules in living cells. *Methods* **98**, 143–149 (2016).

86. Schafer, S. *et al.* Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). in *Current Protocols in Human Genetics* 11.16.1-11.16.14 (John Wiley & Sons, Inc., 2015). doi:10.1002/0471142905.hg1116s87

87. Kornberg, R. D. The molecular basis of eukaryotic transcription. *Proc. Natl. Acad. Sci.* **104**, 12955–12961 (2007).

88. Singh, J. & Padgett, R. A. Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.* **16**, 1128–33 (2009).

89. Danko, C. G. *et al.* Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Mol. Cell* **50**, 212–222 (2013).

90. Goodwin, E. C. & Rottmang, F. M. The 3'-flanking sequence of the bovine growth hormone gene contains novel elements required for efficient and accurate polyadenylation. *J. Biol. Chem.* **267**, 16330–16334 (1992).

91. Arakawa, K. & Tomita, M. The GC skew index: A measure of genomic compositional asymmetry and the degree of replicational selection. *Evol. Bioinforma.* **3**, 159–168 (2007).

92. Merkulov, V. M., Klimova, N. V & Merkulova, T. I. Glucocorticoid Receptor : Translocation from the Cytoplasm to the Nuclei ; Chromatin and Intranuclear Chaperone Cycles. *Russ. J. Genet. Appl. Res.* **6**, 297–306 (2016).

93. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 167–177 (2015).

94. Gromak, N., West, S. & Proudfoot, N. J. Pause Sites Promote Transcriptional Termination of Mammalian RNA Polymerase II. *Mol. Cell. Biol.* **26**, 3986–3996 (2006).

95. Martins, S. B. *et al.* Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3′ end of human genes. *Nat. Struct. Mol. Biol.* **18**, 1115–1123 (2011).

96. de Almeida, S. F., García-Sacristán, A., Custódio, N. & Carmo-Fonseca, M. A link between nuclear RNA surveillance, the human exosome and RNA polymerase II transcriptional termination. *Nucleic Acids Res.* **38**, 8015–8026 (2010).

97. Ruskin, B. & Green, M. R. Role of the 3′ splice site consensus sequence in mammalian pre-mRNA splicing. *Nature* **317**, 732–734 (1985).

98. Clancy, S. RNA Splicing: Introns, Exons and Spliceosome. *Nat. Educ.* **1**, 31 (2008).

99. Gooding, C., Roberts, G. C. & Smith, C. W. J. Role of an inhibitory pyrimidine element and polypyrimidine tract binding protein in repression of a regulated alpha-tropomyosin exon. *Rna* **4**, 85–100 (1998).

100. Bresson, S. & Tollervey, D. Surveillance-ready transcription: nuclear RNA decay as a default fate. *Open Biol.* **8**, 170270 (2018).

101. De La Mata, M. *et al.* A slow RNA polymerase II affects alternative splicing in vivo. *Mol. Cell* **12**, 525–532 (2003).

102. Kadener, S. Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. *EMBO J.* **20**, 5759–5768 (2001).

103. Yanling Zhao, D. *et al.* SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination. *Nature* **529**, 48–53 (2016).

104. Bhatia, V. *et al.* BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature* **511**, 362–365 (2014).

105. Kaneko, S., Rozenblatt-Rosen, O., Meyerson, M. & Manley, J. L. The multifunctional protein p54nrb/PSF recruits the exonuclease XRN2 to facilitate pre-mRNA 3' processing and transcription termination. *Genes & Dev.* **21**, 1779–1789 (2007).

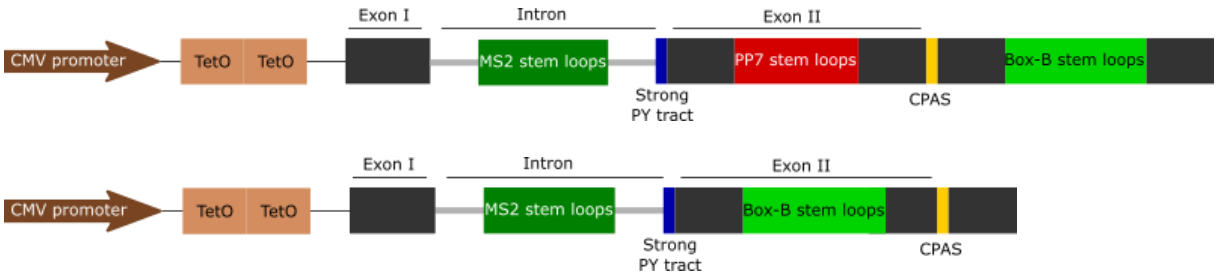
106. Miki, T. S. & Großhans, H. The multifunctional RNase XRN2. *Biochem. Soc. Trans.* **41**, 825–830 (2013).

107. Brody, Y. *et al.* The In Vivo Kinetics of RNA Polymerase II Elongation during Co-Transcriptional Splicing. *PLoS Biol.* **9**, e1000573 (2011).

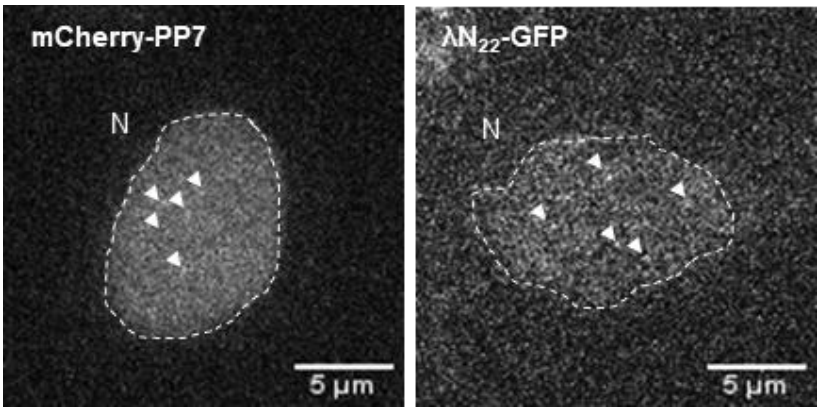
6. Annexes

Annex 1

A



B



C

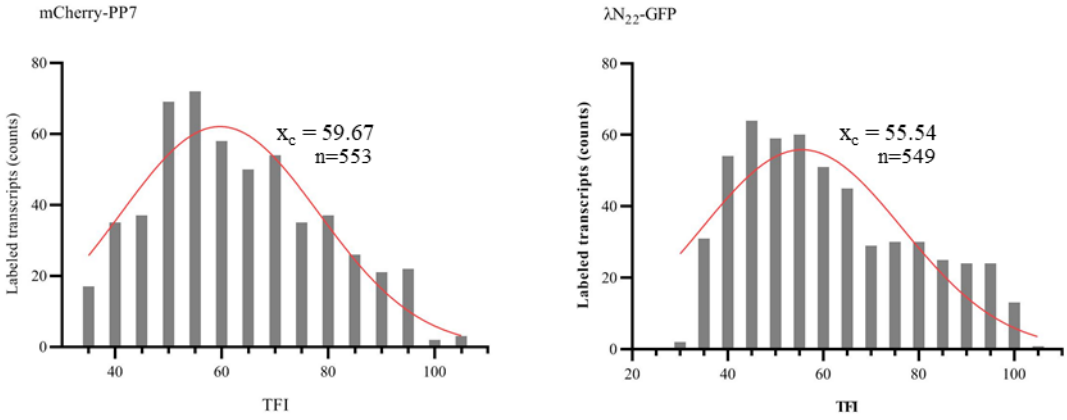
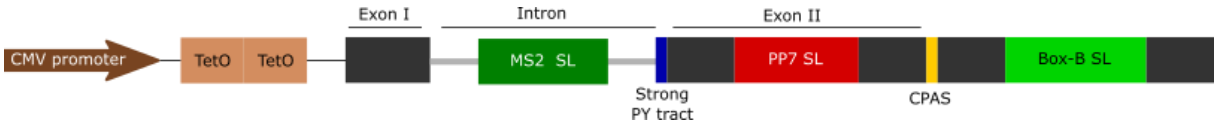
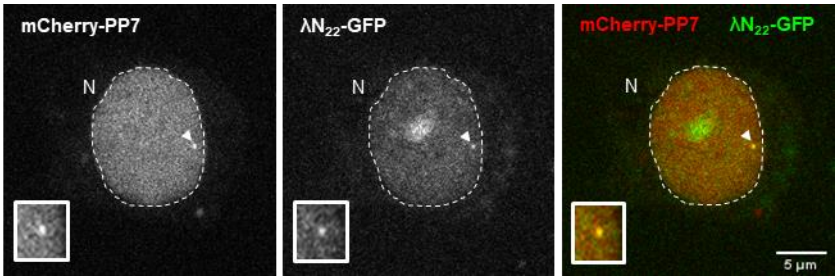


Figure 6.1 Single-molecule calibration measurements. **A** Schematic depiction of the reporter genes used for calibration experiments. The reporter gene on the top was used for calibration of the TFI over background for individual transcripts labeled with mCherry-PP7 and the one on the bottom for calibration of the TFI over background for individual transcripts labeled with λN_{22}-GFP. **B** Microscopy images of diffusing nucleoplasmic reporter gene mRNAs labeled with mCherry-PP7 or λN_{22}-GFP. The nucleus (N) is delimited with a dashed line and the TS of the reporter gene is indicated in each cell with an arrowhead. **C** A gauss fit over histogram plotted TFI measurements determines mean transcript TFI values for the two labeling systems PP7 and λN_{22}.

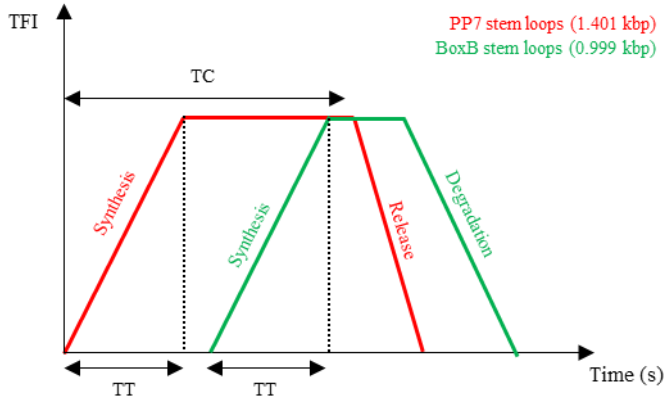
Annex 2



B



C



$$\textbf{RNA Pol II speed} = \frac{\textbf{Lenght of the stem-loops sequence}}{\textbf{Time of transcription}}$$

Figure 6.2 Analysis of live cell spinning-disk confocal images. **A** Scheme of the IgM reporter gene under control of a CMV promoter and two Tet-operator sequences. Binding sites for mCherry-PP7and λN_{22}-GFP were inserted in the exon II and after CPAS, respectively. Bindings sites for MS2 were inserted in the intron. The interaction of fluorescently tagged RNA stem-loop binding proteins with the respective stem-loops allows the visualization of the RNAs. **B** Microscopy images of HEK 293 cells transfected with mCherry-PP7 and λN_{22}-GFP expression plasmids to fluorescently label the PP7 and BoxB stem-loops. The TS is indicated by an arrowhead and the nucleus (N) is limited with a dash line. Corresponding insets depicting the diffraction limited objects in the highest intensity plane of acquired z-stacks are shown below. **C** Schematic depiction of transcription of single RNA transcripts. The time to transcribe the PP7 and BoxB stem-loops corresponds to the time between the increase of the fluorescence signal from the background level to a plateau of intensity. Transcription rates were measured by computing different labelling offset from the PP7 (in red) or BoxB (in green) sequence length within given time intervals for transcription.

Annex 3

Table 6.1 Sequences of the primers used to detect the spliced and unspliced transcripts synthetized from the reporter genes (qPCR)

Transcripts detected	Primer designation	Sequence	Size of the product (bp)
Spliced	Fw	GAATTCTGCAGTCGACGGTAC	270
	Rev	CAGGTCAGGGTGGTCACG	
Unspliced	Fw	AGAGGTGTTTGAGGACACAGG	338
	Rev	CAGGTCAGGGTGGTCACG	
PCNA (housekeeping gene)	Fw	GTTGCAGGCGTAGCAGAGTG	153
	Rev	GGTGGCGGAGTGGCAACAAC	

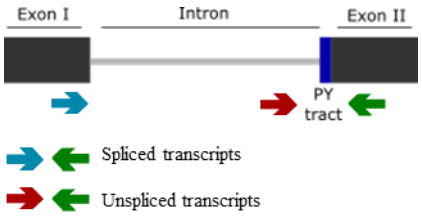


Figure 6.3 Evaluation of reporter gene splicing efficiency. To evaluate the ratio of spliced and unspliced transcripts, for the three different reporter cell lines, we used primer pairs for amplicons between exon I/intron and exon I/exon II for the spliced and unspliced transcripts.